

POSITION DEPENDENT RECOGNITION OF GNN NUCLEOTIDE TRIPLETS BY ZINC FINGERS

5 CROSS-REFERENCES TO RELATED APPLICATIONS

The present application is a continuation-in-part of copending U.S. Patent Application Serial No. 09/535,008, filed March 23, 2000, which application claims the benefit of U.S. provisional applications 60/126,238, filed March 24, 1999, 60/126,239 filed March 24, 1999, 60/146,595 filed July 30, 1999 and 60/146,615 filed July 30, 1999.

10 The present application is also a continuation-in-part of copending U.S. Patent Application Serial No. 09/716,637, filed November 20, 2000. The disclosures of all of the aforementioned applications are hereby incorporated by reference in their entireties for all purposes.

15 BACKGROUND

Zinc finger proteins (ZFPs) are proteins that can bind to DNA in a sequence-specific manner. Zinc fingers were first identified in the transcription factor TFIIIA from the oocytes of the African clawed toad, *Xenopus laevis*. An exemplary motif characterizing one class of these protein (C_2H_2 class) is -Cys-(X)₂₋₄-Cys-(X)₁₂-His-(X)₃₋₅-
20 His (where X is any amino acid) (SEQ. ID. No:1). A single finger domain is about 30 amino acids in length, and several structural studies have demonstrated that it contains an alpha helix containing the two invariant histidine residues and two invariant cysteine residues in a beta turn co-ordinated through zinc. To date, over 10,000 zinc finger sequences have been identified in several thousand known or putative transcription
25 factors. Zinc finger domains are involved not only in DNA-recognition, but also in RNA binding and in protein-protein binding. Current estimates are that this class of molecules will constitute about 2% of all human genes.

The x-ray crystal structure of Zif268, a three-finger domain from a murine transcription factor, has been solved in complex with a cognate DNA sequence and
30 shows that each finger can be superimposed on the next by a periodic rotation. The structure suggests that each finger interacts independently with DNA over 3 base-pair

intervals, with side-chains at positions -1, 2, 3 and 6 on each recognition helix making contacts with their respective DNA triplet subsites. The amino terminus of Zif268 is situated at the 3' end of the DNA strand with which it makes most contacts. Some zinc fingers can bind to a fourth base in a target segment. If the strand with which a zinc

5 finger protein makes most contacts is designated the target strand, some zinc finger proteins bind to a three base triplet in the target strand and a fourth base on the nontarget strand. The fourth base is complementary to the base immediately 3' of the three base subsite.

The structure of the Zif268-DNA complex also suggested that the DNA sequence

10 specificity of a zinc finger protein might be altered by making amino acid substitutions at the four helix positions (-1, 2, 3 and 6) on each of the zinc finger recognition helices. Phage display experiments using zinc finger combinatorial libraries to test this observation were published in a series of papers in 1994 (Rebar et al., *Science* 263, 671-673 (1994); Jamieson et al., *Biochemistry* 33, 5689-5695 (1994); Choo et al, *PNAS* 91,

15 11163-11167 (1994)). Combinatorial libraries were constructed with randomized side-chains in either the first or middle finger of Zif268 and then used to select for an altered Zif268 binding site in which the appropriate DNA sub-site was replaced by an altered DNA triplet. Further, correlation between the nature of introduced mutations and the resulting alteration in binding specificity gave rise to a partial set of substitution rules for

20 design of ZFPs with altered binding specificity.

Greisman & Pabo, *Science* 275, 657-661 (1997) discuss an elaboration of the phage display method in which each finger of a Zif268 was successively randomized and selected for binding to a new triplet sequence. This paper reported selection of ZFPs for a nuclear hormone response element, a p53 target site and a TATA box sequence.

25 A number of papers have reported attempts to produce ZFPs to modulate particular target sites. For example, Choo et al., *Nature* 372, 645 (1994), report an attempt to design a ZFP that would repress expression of a bcr-abl oncogene. The target segment to which the ZFPs would bind was a nine base sequence 5'GCA GAA GCC3' chosen to overlap the junction created by a specific oncogenic translocation fusing the

30 genes encoding bcr and abl. The intention was that a ZFP specific to this target site would bind to the oncogene without binding to abl or bcr component genes. The authors

used phage display to screen a mini-library of variant ZFPs for binding to this target segment. A variant ZFP thus isolated was then reported to repress expression of a stably transfected bcr-able construct in a cell line.

Pomerantz et al., *Science* 267, 93-96 (1995) reported an attempt to design a novel DNA binding protein by fusing two fingers from Zif268 with a homeodomain from Oct-1. The hybrid protein was then fused with a transcriptional activator for expression as a chimeric protein. The chimeric protein was reported to bind a target site representing a hybrid of the subsites of its two components. The authors then constructed a reporter vector containing a luciferase gene operably linked to a promoter and a hybrid site for the chimeric DNA binding protein in proximity to the promoter. The authors reported that their chimeric DNA binding protein could activate expression of the luciferase gene.

Liu et al., *PNAS* 94, 5525-5530 (1997) report forming a composite zinc finger protein by using a peptide spacer to link two component zinc finger proteins each having three fingers. The composite protein was then further linked to transcriptional activation domain. It was reported that the resulting chimeric protein bound to a target site formed from the target segments bound by the two component zinc finger proteins. It was further reported that the chimeric zinc finger protein could activate transcription of a reporter gene when its target site was inserted into a reporter plasmid in proximity to a promoter operably linked to the reporter.

Choo et al., WO 98/53058, WO98/53059, and WO 98/53060 (1998) discuss selection of zinc finger proteins to bind to a target site within the HIV Tat gene. Choo et al. also discuss selection of a zinc finger protein to bind to a target site encompassing a site of a common mutation in the oncogene ras. The target site within ras was thus constrained by the position of the mutation.

Previously-disclosed methods for the design of sequence-specific zinc finger proteins have often been based on modularity of individual zinc fingers; *i.e.*, the ability of a zinc finger to recognize the same target subsite regardless of the location of the finger in a multi-finger protein. Although, in many instances, a zinc finger retains the same sequence specificity regardless of its location within a multi-finger protein; in certain cases, the sequence specificity of a zinc finger depends on its position. For example, it is possible for a finger to recognize a particular triplet sequence when it is

present as finger 1 of a three-finger protein, but to recognize a different triplet sequence when present as finger 2 of a three-finger protein.

Attempts to address situations in which a zinc finger behaves in a non-modular fashion (*i.e.*, its sequence specificity depends upon its location in a multi-finger protein) have, to date, involved strategies employing randomization of key binding residues in multiple adjacent zinc fingers, followed by selection. *See*, for example, Isalan *et al.* (2001) *Nature Biotechnol.* **19**:656-660. However, methods for rational design of polypeptides containing non-modular zinc fingers have not heretofore been described.

SUMMARY

The present disclosure provides compositions comprising and methods involving position dependent recognition of GNN nucleotide triplets by zinc fingers.

Thus, provided herein is a zinc finger protein that binds to a target site, said zinc finger protein comprising a first (F1), a second (F2), and a third (F3) zinc finger, ordered F1, F2, F3 from N-terminus to C-terminus, said target site comprising, in 3' to 5' direction, a first (S1), a second (S2), and a third (S3) target subsite, each target subsite having the nucleotide sequence GNN, wherein if S1 comprises GAA, F1 comprises the amino acid sequence QRSNLVR; if S2 comprises GAA, F2 comprises the amino acid sequence QSGNLAR; if S3 comprises GAA, F3 comprises the amino acid sequence QSGNLAR; if S1 comprises GAG, F1 comprises the amino acid sequence RSDNLAR; if S2 comprises GAG, F2 comprises the amino acid sequence RSDNLAR; if S3 comprises GAG, F3 comprises the amino acid sequence RSDNLTR; if S1 comprises GAC, F1 comprises the amino acid sequence DRSNLTR; if S2 comprises GAC, F2 comprises the amino acid sequence DRSNLTR; if S3 comprises GAC, F3 comprises the amino acid sequence DRSNLTR; if S1 comprises GAT, F1 comprises the amino acid sequence QSSNLAR; if S2 comprises GAT, F2 comprises the amino acid sequence TSGNLVR; if S3 comprises GAT, F3 comprises the amino acid sequence TSANLSR; if S1 comprises GGA, F1 comprises the amino acid sequence QSGHLAR; if S2 comprises GGA, F2 comprises the amino acid sequence QSGHLQR; if S3 comprises GGA, F3 comprises the amino acid sequence QSGHLQR; if S1 comprises GGG, F1 comprises the amino acid sequence RSDHLAR; if S2 comprises GGG, F2 comprises the amino acid sequence

RSDHLSR; if S3 comprises GGG, F3 comprises the amino acid sequence RSDHLSR; if
 S1 comprises GGC, F1 comprises the amino acid sequence DRSHLRT; if S2 comprises
 GGC, F2 comprises the amino acid sequence DRSHLAR; if S1 comprises GGT, F1
 comprises the amino acid sequence QSSHLTR; if S2 comprises GGT, F2 comprises the
 5 amino acid sequence TSGHLSR; if S3 comprises GGT, F3 comprises the amino acid
 sequence TSGHLVR; if S1 comprises GCA, F1 comprises the amino acid sequence
 QSGSLTR; if S2 comprises GCA, F2 comprises QSGDLTR; if S3 comprises GCA, F3
 comprises QSGDLTR; if S1 comprises GCG, F1 comprises the amino acid sequence
 RSDDLTR; if S2 comprises GCG, F2 comprises the amino acid sequence RSDDLQR; if
 10 S3 comprises GCG, F3 comprises the amino acid sequence RSDDLTR; if S1 comprises
 GCC, F1 comprises the amino acid sequence ERGTLAR; if S2 comprises GCC, F2
 comprises the amino acid sequence DRSDLTR; if S3 comprises GCC, F3 comprises the
 amino acid sequence DRSDLTR; if S1 comprises GCT, F1 comprises the amino acid
 sequence QSSDLTR; if S2 comprises GCT, F2 comprises the amino acid sequence
 15 QSSDLTR; if S3 comprises GCT, F3 comprises the amino acid sequence QSSDLQR; if
 S1 comprises GTA, F1 comprises the amino acid sequence QSGALTR; if S2 comprises
 GTA, F2 comprises the amino acid sequence QSGALAR; if S1 comprises GTG, F1
 comprises the amino acid sequence RSDALTR; if S2 comprises GTG, F2 comprises the
 amino acid sequence RSDALSR; if S3 comprises GTG, F3 comprises the amino acid
 20 sequence RSDALTR; if S1 comprises GTC, F1 comprises the amino acid sequence
 DRSLAR; if S2 comprises GTC, F2 comprises the amino acid sequence DRSLAR;
 and if S3 comprises GTC, F3 comprises the amino acid sequence DRSLAR.

Also provided are methods of designing a zinc finger protein comprising a first
 (F1), a second (F2), and a third (F3) zinc finger, ordered F1, F2, F3 from N-terminus to
 25 C-terminus that binds to a target site comprising, in 3' to 5' direction, a first (S1), a
 second (S2), and a third (S3) target subsite, each target subsite having the nucleotide
 sequence GNN, the method comprising the steps of (a) selecting the F1 zinc finger such
 that it binds to the S1 target subsite, wherein if S1 comprises GAA, F1 comprises the
 amino acid sequence QRSNLVR; if S1 comprises GAG, F1 comprises the amino acid
 30 sequence RSDNLAR; if S1 comprises GAC, F1 comprises the amino acid sequence
 DRSNLTR; if S1 comprises GAT, F1 comprises the amino acid sequence QSSNLAR; if

S1 comprises GGA, F1 comprises the amino acid sequence QSGHLAR; if S1 comprises
 GGG, F1 comprises the amino acid sequence RSDHLAR; if S1 comprises GGC, F1
 comprises the amino acid sequence DRSHLRT; if S1 comprises GGT, F1 comprises the
 amino acid sequence QSSHLTR; if S1 comprises GCA, F1 comprises QSGSLTR; if S1
 5 comprises GCG, F1 comprises RSDDLTR; if S2 comprises GCG, F2 comprises
 RSDDLQR; if S1 comprises GCC, F1 comprises ERGTLAR; if S1 comprises GCT, F1
 comprises the amino acid sequence QSSDLTR; if S1 comprises GTA, F1 comprises the
 amino acid sequence QSGALTR; if S1 comprises GTG, F1 comprises the amino acid
 sequence RSDALTR; if S1 comprises GTC, F1 comprises the amino acid sequence
 10 DRSLAR; (b) selecting the F2 zinc finger such that it binds to the S2 target subsite,
 wherein S2 comprises GAA, F2 comprises the amino acid sequence QSGNLAR; if S2
 comprises GAG, F2 comprises the amino acid sequence RSDNLAR; if S2 comprises
 GAC, F2 comprises the amino acid sequence DRSNLTR; if S2 comprises GAT, F2
 comprises the amino acid sequence TSGNLVR; if S2 comprises GGA, F2 comprises the
 15 amino acid sequence QSGHLQR; if S2 comprises GGG, F2 comprises the amino acid
 sequence RSDHLSR; if S2 comprises GGC, F2 comprises the amino acid sequence
 DRSHLAR; if S2 comprises GGT, F2 comprises the amino acid sequence TSGHLSR; if
 S2 comprises GCA, F2 comprises the amino acid sequence QSGDLTR; if S2 comprises
 GCC, F2 comprises the amino acid sequence DRSDLTR; if S2 comprises GCT, F2
 20 comprises the amino acid sequence QSSDLTR; if S2 comprises GTA, F2 comprises the
 amino acid sequence QSGALAR; if S2 comprises GTG, F2 comprises the amino acid
 sequence RSDALSR; if S2 comprises GTC, F2 comprises the amino acid sequence
 DRSLAR; and (c) selecting the F3 zinc finger such that it binds to the S3 target subsite,
 wherein if S3 comprises GAA, F3 comprises the amino acid sequence QSGNLAR; if S3
 25 comprises GAG, F3 comprises the amino acid sequence RSDNLTR; if S3 comprises
 GAC, F3 comprises the amino acid sequence DRSNLTR; if S3 comprises GAT, F3
 comprises the amino acid sequence TSANLSR; if S3 comprises GGA, F3 comprises the
 amino acid sequence QSGHLQR; if S3 comprises GGG, F3 comprises RSDHLSR; if S3
 comprises GGT, F3 comprises the amino acid sequence TSGHLVR; if S3 comprises
 30 GCA, F3 comprises the amino acid sequence QSGDLTR; if S3 comprises GCG, F3
 comprises the amino acid sequence RSDDLTR; if S3 comprises GCC, F3 comprises the

amino acid sequence DRSDLTR; if S3 comprises GCT, F3 comprises the amino acid sequence QSSDLQR; if S3 comprises GTG, F3 comprises RSDALTR; and if S3 comprises GTC, F3 comprises the amino acid sequence DRSALAR;

thereby designing a zinc finger protein that binds to a target site.

- 5 In certain embodiments of the zinc finger proteins and methods described herein, S1 comprises GAA and F1 comprises the amino acid sequence QRSNLVR. In other embodiments, S2 comprises GAA and F2 comprises the amino acid sequence QSGNLAR. In other embodiments, S3 comprises GAA and F3 comprises the amino acid sequence QSGNLAR. In other embodiments, S1 comprises GAG and F1 comprises the amino acid sequence RSDNLAR. In other embodiments, S2 comprises GAG and F2 comprises the amino acid sequence RSDNLAR. In other embodiments, S3 comprises GAG and F3 comprises the amino acid sequence RSDNLTR. In other embodiments, S1 comprises GAC and F1 comprises the amino acid sequence DRSNLTR. In other embodiments, S2 comprises GAC and F2 comprises the amino acid sequence DRSNLTR. In other embodiments, S3 comprises GAC and F3 comprises the amino acid sequence DRSNLTR. In other embodiments, S1 comprises GAT and F1 comprises the amino acid sequence QSSNLAR. In other embodiments, S2 comprises GAT and F2 comprises the amino acid sequence TSGNLVR. In other embodiments, S3 comprises GAT and F3 comprises the amino acid sequence TSANLSR. In other embodiments, S1 comprises GGA and F1 comprises the amino acid sequence QSGHLAR. In other embodiments, S2 comprises GGA and F2 comprises the amino acid sequence QSGHLQR. In other embodiments, S3 comprises GGA and F3 comprises the amino acid sequence QSGHLQR. In other embodiments, S1 comprises GGG and F1 comprises the amino acid sequence RSDHLAR. In other embodiments, S2 comprises GGG and F2 comprises the amino acid sequence RSDHLSR. In other embodiments, S3 comprises GGG and F3 comprises the amino acid sequence RSDHLSR. In other embodiments, S1 comprises GGC and F1 comprises the amino acid sequence DRSHLTR. In other embodiments, S2 comprises GGC and F2 comprises the amino acid sequence DRSHLAR. In other embodiments, S1 comprises GGT and F1 comprises the amino acid sequence QSSHLTR. In other embodiments, S2 comprises GGT and F2 comprises the amino acid sequence TSGHLSR. In other embodiments, S3 comprises GGT and F3

- comprises the amino acid sequence TSGHLVR. In other embodiments, S1 comprises GCA and F1 comprises the amino acid sequence QSGSLTR. In other embodiments, S2 comprises GCA and F2 comprises the amino acid sequence QSGDLTR. In other embodiments, S3 comprises GCA and F3 comprises the amino acid sequence
- 5 QSGDLTR. In other embodiments, S1 comprises GCG and F1 comprises the amino acid sequence RSDDLTR. In other embodiments, S2 comprises GCG and F2 comprises the amino acid sequence RSDDLQR. In other embodiments, S3 comprises GCG and F3 comprises the amino acid sequence RSDDLTR. In other embodiments, S1 comprises GCC and F1 comprises the amino acid sequence ERGTLAR. In other embodiments, S2
- 10 comprises GCC and F2 comprises the amino acid sequence DRSDLTR. In other embodiments, S3 comprises GCC and F3 comprises the amino acid sequence DRSDLTR. In other embodiments, S1 comprises GCT and F1 comprises the amino acid sequence QSSDLTR. In other embodiments, S2 comprises GCT and F2 comprises the amino acid sequence QSSDLTR. In other embodiments, S3 comprises GCT and F3 comprises the amino acid sequence QSSDLQR. In other embodiments, S1 comprises GTA and F1
- 15 comprises the amino acid sequence QSGALTR. In other embodiments, S2 comprises GTA and F2 comprises the amino acid sequence QSGALAR. In other embodiments, S1 comprises GTG and F1 comprises the amino acid sequence RSDALTR. In other embodiments, S2 comprises GTG and F2 comprises the amino acid sequence RSDALSR.
- 20 In other embodiments, S3 comprises GTG and F3 comprises the amino acid sequence RSDALTR. In other embodiments, S1 comprises GTC and F1 comprises the amino acid sequence DRSALAR. In other embodiments, S2 comprises GTC and F2 comprises the amino acid sequence DRSALAR. In other embodiments, S3 comprises GTC and F3 comprises the amino acid sequence DRSALAR.
- 25 Also provided are polypeptides comprising any of zinc finger proteins described herein. In certain embodiments, the polypeptide further comprises at least one functional domain. Also provided are polynucleotides encoding any of the polypeptides described herein. Thus, also provided are nucleic acid encoding zinc fingers, including all of the zinc fingers described above.

Also provided are segments of a zinc finger comprising a sequence of seven contiguous amino acids as shown herein. Also provided are nucleic acids encoding any of these segments and zinc fingers comprising the same.

Also provided are zinc finger proteins comprising first, second and third zinc
5 fingers. The first, second and third zinc fingers comprise respectively first, second and third segments of seven contiguous amino acids as shown herein. Also provided are nucleic acids encoding such zinc finger proteins.

BRIEF DESCRIPTION OF THE DRAWINGS

10 **Figure 1** shows results of site selection analysis of two representative zinc finger proteins (leftmost 4 columns) and measurements of binding affinity for each of these proteins to their intended target sequences and to variant target sequences. (rightmost 3 columns). Analysis of ZFP1 is shown in the upper portion of the figure and analysis of ZFP2 is shown in the lower portion of the figure. For the site selection analyses, the
15 amino acid sequences of residues -1 through +6 of the recognition helix of each of the three component zinc fingers (F3, F2 and F1) are shown across the top row; the intended target sequence (divided into finger-specific target subsites) is shown across the second row, and a summary of the sequences bound is shown in the third row. Data for F3 is shown in the second column, data for F2 is shown in the third column, and data for F1 is
20 shown in the third column.

For the binding affinity analyses, the designed target sequence for each ZFP ("cognate") and two related sequences ("Mt") are shown (column 6), along with the K_d for binding of the ZFP to each of these sequences (column 7).

25 **Figure 2** shows amino acid sequences of zinc finger recognition regions (amino acids -1 through +6 of the recognition helix) that bind to each of the 16 GNN triplet subsites. Three amino acid sequences are shown for each trinucleotide subsite; these correspond to optimal amino acid sequences for recognition of the subsite from each of the three positions (finger 1, F1; finger 2, F2; or finger 3, F3) in a three-finger zinc finger protein. Amino acid sequences are from N-terminal to C-terminal; nucleotide sequences
30 are from 5' to 3'.

Also shown are site selection results for each of the 48 position-dependent GNN-recognizing zinc fingers. These show the number of times a particular nucleotide was present, at a given position, in a collection of oligonucleotide sequences bound by the finger. For example, out of 15 oligonucleotides bound by a zinc finger protein with the amino acid sequence QSGHLAR present at the finger 1 (F1) position, 15 contained a G in the 5'-most position of the subsite, 15 contained a G in the middle position of the subsite, while, at the 3'-most position of the subsite, 10 contained an A, 3 contained a G and 2 contained a T. Accordingly, this particular amino acid sequence is optimal for binding a GGA triplet from the F1 position.

Figures 3A, 3B and 3C show site selection data indicating positional dependence of GCA-, GAT- and GGT-binding zinc fingers. The first and fourth (where applicable) rows of each figure show portions of the amino acid sequence of a designed zinc finger protein. Amino acid residues -1 through +6 of each α -helix are listed from left to right. The second and fifth (where applicable) rows show the target sequence, divided into three triplet subsites, one for each finger of the protein shown in the first and fourth (where applicable) rows, respectively. The third and sixth (where applicable) rows show the distribution of nucleotides in the oligonucleotides obtained by site selection with the proteins shown in the first and fourth (where applicable) rows, respectively. Figure 3A shows data for fingers designed to bind GCA; Figure 3B shows data for fingers designed to bind GAT; Figure 3C shows data for fingers designed to bind GGT.

Figures 4A and 4B show properties of the engineered ZFP EP2C. Figure 4A shows site selection data. The first row provides the amino acid sequences of residues -1 through +6 of the recognition helices for each of the three zinc fingers of the EP2C protein. The second row shows the target sequence (5' to 3'); with the distribution of nucleotides in the oligonucleotides obtained by site selection indicated below the target sequence.

Figure 4B shows *in vitro* and *in vivo* assays for the binding specificity of EP2C. The first three columns show *in vitro* measurements of binding affinity of EP2C to its intended target sequence and several related sequences. The first column gives the name of each sequence (2C0 is the intended target sequence, compare to Figure 4A). The second column shows the nucleotide sequence of various target sequences, with

differences from the intended target sequence (2C0) highlighted. The third column shows the K_d (in nM) for binding of EP2C to each of the target sequences. K_d s were determined by gel shift assays, using 2-fold dilution series of EP2C. The right side of the figure (fourth column and bar graph) shows relative luciferase activities (normalized to β -galactosidase levels) in stable cell lines in which expression of EP2C is inducible. Cells were co-transfected with a vector containing a luciferase coding region under the transcriptional control of the target sequence shown in the same row of the figure, and a control vector encoding β -galactosidase. Luciferase and β -galactosidase levels were measured after induction of EP2C expression. Triplicate samples were assayed and the standard deviations are shown in the bar graph. pGL3 is a luciferase-encoding vector lacking EP2C target sequences. 3B is another negative control, in which luciferase expression is under transcriptional control of sequences (3B) unrelated to the EP2C target sequence.

DEFINITIONS

A zinc finger DNA binding protein is a protein or segment within a larger protein that binds DNA in a sequence-specific manner as a result of stabilization of protein structure through coordination of a zinc ion. The term zinc finger DNA binding protein is often abbreviated as zinc finger protein or ZFP.

Zinc finger proteins can be engineered to recognize a selected target sequence in a nucleic acid. Any method known in the art or disclosed herein can be used to construct an engineered zinc finger protein or a nucleic acid encoding an engineered zinc finger protein. These include, but are not limited to, rational design, selection methods (*e.g.*, phage display) random mutagenesis, combinatorial libraries, computer design, affinity selection, use of databases matching zinc finger amino acid sequences with target subsite nucleotide sequences, cloning from cDNA and/or genomic libraries, and synthetic constructions. An engineered zinc finger protein can comprise a new combination of naturally-occurring zinc finger sequences. Methods for engineering zinc finger proteins are disclosed in co-owned WO 00/41566 and WO 00/42219; as well as in WO 98/53057; WO 98/53058; WO 98/53059 and WO 98/53060; the disclosures of which are hereby incorporated by reference in their entireties. Methods for identifying preferred target

sequences, and for engineering zinc finger proteins to bind to such preferred target sequences, are disclosed in co-owned WO 00/42219.

5 A designed zinc finger protein is a protein not occurring in nature whose design/composition results principally from rational criteria. Rational criteria for design include application of substitution rules and computerized algorithms for processing information in a database storing information of existing ZFP designs and binding data.

A selected zinc finger protein is a protein not found in nature whose production results primarily from an empirical process such as phage display.

10 The term naturally-occurring is used to describe an object that can be found in nature as distinct from being artificially produced by man. For example, a polypeptide or polynucleotide sequence that is present in an organism (including viruses) that can be isolated from a source in nature and which has not been intentionally modified by man in the laboratory is naturally-occurring. Generally, the term naturally-occurring refers to an object as present in a non-pathological (undiseased) individual, such as would be typical
15 for the species.

A nucleic acid is operably linked when it is placed into a functional relationship with another nucleic acid sequence. For instance, a promoter or enhancer is operably linked to a coding sequence if it increases the transcription of the coding sequence. Operably linked means that the DNA sequences being linked are typically contiguous
20 and, where necessary to join two protein coding regions, contiguous and in reading frame. However, since enhancers generally function when separated from the promoter by up to several kilobases or more and intronic sequences may be of variable lengths, some polynucleotide elements may be operably linked but not contiguous.

A specific binding affinity between, for example, a ZFP and a specific target site
25 means a binding affinity of at least $1 \times 10^6 \text{ M}^{-1}$.

The terms "modulating expression" "inhibiting expression" and "activating expression" of a gene refer to the ability of a zinc finger protein to activate or inhibit transcription of a gene. Activation includes prevention of subsequent transcriptional inhibition (i.e., prevention of repression of gene expression) and inhibition includes
30 prevention of subsequent transcriptional activation (i.e., prevention of gene activation). Modulation can be assayed by determining any parameter that is indirectly or directly

affected by the expression of the target gene. Such parameters include, e.g., changes in RNA or protein levels, changes in protein activity, changes in product levels, changes in downstream gene expression, changes in reporter gene transcription (luciferase, CAT, beta-galactosidase, GFP (see, e.g., Mistili & Spector, *Nature Biotechnology* 15:961-964 (1997))); changes in signal transduction, phosphorylation and dephosphorylation, receptor-ligand interactions, second messenger concentrations (e.g., cGMP, cAMP, IP3, and Ca²⁺), cell growth, neovascularization, *in vitro*, *in vivo*, and *ex vivo*. Such functional effects can be measured by any means known to those skilled in the art, e.g., measurement of RNA or protein levels, measurement of RNA stability, identification of downstream or reporter gene expression, e.g., via chemiluminescence, fluorescence, colorimetric reactions, antibody binding, inducible markers, ligand binding assays; changes in intracellular second messengers such as cGMP and inositol triphosphate (IP3); changes in intracellular calcium levels; cytokine release, and the like.

A "regulatory domain" refers to a protein or a protein subsequence that has transcriptional modulation activity. Typically, a regulatory domain is covalently or non-covalently linked to a ZFP to modulate transcription. Alternatively, a ZFP can act alone, without a regulatory domain, or with multiple regulatory domains to modulate transcription.

A D-able subsite within a target site has the motif 5'NNGK3'. A target site containing one or more such motifs is sometimes described as a D-able target site. A zinc finger appropriately designed to bind to a D-able subsite is sometimes referred to as a D-able finger. Likewise a zinc finger protein containing at least one finger designed or selected to bind to a target site including at least one D-able subsite is sometimes referred to as a D-able zinc finger protein.

DETAILED DESCRIPTION

I. General

Tables 1-5 list a collection of nonnaturally occurring zinc finger protein sequences and their corresponding target sites. The first column of each table is an internal reference number. The second column lists a 9 or 10 base target site bound by a three-finger zinc finger protein, with the target sites listed in 5' to 3' orientation. The

third column provides SEQ ID NOs for the target site sequences listed in column 2. The fourth, sixth and eighth columns list amino acid residues from the first, second and third fingers, respectively, of a zinc finger protein which recognizes the target sequence listed in the second column. For each finger, seven amino acids, occupying positions -1 to +6 of the finger, are listed. The numbering convention for zinc fingers is defined below. Columns 5, 7 and 9 provide SEQ ID NOs for the amino acid sequences listed in columns 4, 6 and 8, respectively. The final column of each table lists the binding affinity (*i.e.*, the K_d in nM) of the zinc finger protein for its target site. Binding affinities are measured as described below.

- Each finger binds to a triplet of bases within a corresponding target sequence. The first finger binds to the first triplet starting from the 3' end of a target site, the second finger binds to the second triplet, and the third finger binds the third (*i.e.*, the 5'-most) triplet of the target sequence. For example, the RSDSLTS finger (SEQ ID NO: 646) of SBS# 201 (Table 2) binds to 5'TTG3', the ERSTLTR finger (SEQ ID NO: 851) binds to 5'GCC3' and the QRADLRR finger (SEQ ID NO: 1056) binds to 5'GCA3'.

- Table 6 lists a collection of consensus sequences for zinc fingers and the target sites bound by such sequences. Conventional one letter amino acid codes are used to designate amino acids occupying consensus positions. The symbol "X" designates a nonconsensus position that can in principle be occupied by any amino acid. In most zinc fingers of the C_2H_2 type, binding specificity is principally conferred by residues -1, +2, +3 and +6. Accordingly, consensus sequence determining binding specificity typically include at least these residues. Consensus sequences are useful for designing zinc fingers to bind to a given target sequence. Residues occupying other positions can be selected based on sequences in Tables 1-5, or other known zinc finger sequences. Alternatively, these positions can be randomized with a plurality of candidate amino acids and screened against one or more target sequences to refine binding specificity or improve binding specificity. In general, the same consensus sequence can be used for design of a zinc finger regardless of the relative position of that finger in a multi-finger zinc finger protein. For example, the sequence RXDNXXR can be used to design a N-terminal, central or C-terminal finger of three finger protein. However, some consensus sequences are most suitable for designing a zinc finger to occupy a particular position in a multi-

finger protein. For example, the consensus sequence RXDHXXQ is most suitable for designing a C-terminal finger of a three-finger protein.

II. Characteristics of Zinc Finger Proteins

- 5 Zinc finger proteins are formed from zinc finger components. For example, zinc finger proteins can have one to thirty-seven fingers, commonly having 2, 3, 4, 5 or 6 fingers. A zinc finger protein recognizes and binds to a target site (sometimes referred to as a target segment) that represents a relatively small subsequence within a target gene. Each component finger of a zinc finger protein can bind to a subsite within the target site.
- 10 The subsite includes a triplet of three contiguous bases all on the same strand (sometimes referred to as the target strand). The subsite may or may not also include a fourth base on the opposite strand that is the complement of the base immediately 3' of the three contiguous bases on the target strand. In many zinc finger proteins, a zinc finger binds to its triplet subsite substantially independently of other fingers in the same zinc finger
- 15 protein. Accordingly, the binding specificity of zinc finger protein containing multiple fingers is usually approximately the aggregate of the specificities of its component fingers. For example, if a zinc finger protein is formed from first, second and third fingers that individually bind to triplets XXX, YYY, and ZZZ, the binding specificity of the zinc finger protein is 3'XXX YYY ZZZ5'.
- 20 The relative order of fingers in a zinc finger protein from N-terminal to C-terminal determines the relative order of triplets in the 3' to 5' direction in the target. For example, if a zinc finger protein comprises from N-terminal to C-terminal first, second and third fingers that individually bind, respectively, to triplets 5' GAC3', 5'GTA3' and 5'GGC3' then the zinc finger protein binds to the target segment
- 25 3'CAGATGCGG5'. If the zinc finger protein comprises the fingers in another order, for example, second finger, first finger, third finger, then the zinc finger protein binds to a target segment comprising a different permutation of triplets, in this example, 3'ATGCAGCGG5' (see Berg & Shi, *Science* 271, 1081-1086 (1996)). The assessment of binding properties of a zinc finger protein as the aggregate of its component fingers
- 30 may, in some cases, be influenced by context-dependent interactions of multiple fingers binding in the same protein.

Two or more zinc finger proteins can be linked to have a target specificity that is the aggregate of that of the component zinc finger proteins (see e.g., Kim & Pabo, *PNAS* 95, 2812-2817 (1998)). For example, a first zinc finger protein having first, second and third component fingers that respectively bind to XXX, YYY and ZZZ can be linked to a second zinc finger protein having first, second and third component fingers with binding specificities, AAA, BBB and CCC. The binding specificity of the combined first and second proteins is thus 3'XXXYYYZZZ__AAABBBCCC5', where the underline indicates a short intervening region (typically 0-5 bases of any type). In this situation, the target site can be viewed as comprising two target segments separated by an intervening segment.

Linkage can be accomplished using any of the following peptide linkers. T G E K P; (SEQ. ID. No:2) (Liu et al., 1997, supra.); (G4S)_n (SEQ. ID. No:3) (Kim et al., *PNAS* 93, 1156-1160 (1996.); GGRRGGGS; (SEQ. ID. No:4) LRQRDGERP; (SEQ. ID. No:5) LRQKDGGGSERP; (SEQ. ID. No:6) LRQKD(G3S)₂ ERP (SEQ. ID. No:7) Alternatively, flexible linkers can be rationally designed using computer programs capable of modeling both DNA-binding sites and the peptides themselves or by phage display methods. In a further variation, noncovalent linkage can be achieved by fusing two zinc finger proteins with domains promoting heterodimer formation of the two zinc finger proteins. For example, one zinc finger protein can be fused with fos and the other with jun (see Barbas et al., WO 95/119431).

Linkage of two zinc finger proteins is advantageous for conferring a unique binding specificity within a mammalian genome. A typical mammalian diploid genome consists of 3×10^9 bp. Assuming that the four nucleotides A, C, G, and T are randomly distributed, a given 9 bp sequence is present ~23,000 times. Thus a ZFP recognizing a 9 bp target with absolute specificity would have the potential to bind to ~23,000 sites within the genome. An 18 bp sequence is present once in 3.4×10^{10} bp, or about once in a random DNA sequence whose complexity is ten times that of a mammalian genome.

A component finger of zinc finger protein typically contains about 30 amino acids and has the following motif (N-C) :

(SEQ. ID. No:8)
Cys- (X)₂₋₄-Cys-X.X.X.X.X.X.X.X.X.X.X.X.X-**His**- (X)₃₋₅-His

-1 1 2 3 4 5 6 7

The two invariant histidine residues and two invariant cysteine residues in a single beta turn are co-ordinated through zinc (see, e.g., Berg & Shi, *Science* 271, 1081-1085 (1996)). The above motif shows a numbering convention that is standard in the field for the region of a zinc finger conferring binding specificity. The amino acid on the left (N-terminal side) of the first invariant His residues is assigned the number +6, and other amino acids further to the left are assigned successively decreasing numbers. The alpha helix begins at residue 1 and extends to the residue following the second conserved histidine. The entire helix is therefore of variable length, between 11 and 13 residues.

The process of designing or selecting a nonnaturally occurring or variant ZFP typically starts with a natural ZFP as a source of framework residues. The process of design or selection serves to define nonconserved positions (i.e., positions -1 to +6) so as to confer a desired binding specificity. One suitable ZFP is the DNA binding domain of the mouse transcription factor Zif268. The DNA binding domain of this protein has the amino acid sequence:

YACPVESCDRRFSRSDDELTRHIRIHTGQKP (F1) (SEQ. ID No:9)
FQCRICMRNFSRSDHLTTHIRTHTGEKP (F2) (SEQ. ID. No:10)
FACDICGRKFARSDERKRHTKIHLRQK (F3) SEQ. ID. No:11)
and binds to a target 5' GCG TGG GCG 3' (SEQ ID No:12).

Another suitable natural zinc finger protein as a source of framework residues is Sp-1. The Sp-1 sequence used for construction of zinc finger proteins corresponds to amino acids 531 to 624 in the Sp-1 transcription factor. This sequence is 94 amino acids in length. The amino acid sequence of Sp-1 is as follows:

PGKKKQHICHIQGCGKVYGKTSHLRAHLRWHTGERP

FMCTWSYCGKRFTRSDDELQRHKRTHHTGEKK

FACPECPKRFMRSDHLSKHIKTHQNKKG (SEQ. ID. No:13)

Sp-1 binds to a target site 5'GGG GCG GGG3' (SEQ ID No: 14).

An alternate form of Sp-1, an Sp-1 consensus sequence, has the following amino acid sequence:

meklmngsgd

PGKKKQHACPECGKSFSKSSHLRAHQRTHTGERP

YKCPECGKSFSRSDELQRHQRTHTGEKP

YKCPECGKSFSRSDHLSKHQRT HQNKKG (SEQ. ID. No:15) (lower case letters are a leader sequence from Shi & Berg, *Chemistry and Biology* 1, 83-89. (1995). The optimal binding sequence for the Sp-1 consensus sequence is 5'GGGGCGGGG3' (SEQ ID No:

5 16) . Other suitable ZFPs are described below.

There are a number of substitution rules that assist rational design of some zinc finger proteins (see Desjarlais & Berg, *PNAS* 90, 2256-2260 (1993); Choo & Klug, *PNAS* 91, 11163-11167 (1994); Desjarlais & Berg, *PNAS* 89, 7345-7349 (1992); Jamieson et al., supra; Choo et al., WO 98/53057, WO 98/53058; WO 98/53059; WO 98/53060).

10 Many of these rules are supported by site-directed mutagenesis of the three-finger domain of the ubiquitous transcription factor, Sp-1 (Desjarlais and Berg, 1992; 1993). One of these rules is that a 5' G in a DNA triplet can be bound by a zinc finger incorporating arginine at position 6 of the recognition helix. Another substitution rule is that a G in the middle of a subsite can be recognized by including a histidine residue at position 3 of a
 15 zinc finger. A further substitution rule is that asparagine can be incorporated to recognize A in the middle of triplet, aspartic acid, glutamic acid, serine or threonine can be incorporated to recognize C in the middle of triplet, and amino acids with small side chains such as alanine can be incorporated to recognize T in the middle of triplet. A further substitution rule is that the 3' base of triplet subsite can be recognized by
 20 incorporating the following amino acids at position -1 of the recognition helix: arginine to recognize G, glutamine to recognize A, glutamic acid (or aspartic acid) to recognize C, and threonine to recognize T. Although these substitution rules are useful in designing zinc finger proteins they do not take into account all possible target sites. Furthermore, the assumption underlying the rules, namely that a particular amino acid in a zinc finger
 25 is responsible for binding to a particular base in a subsite is only approximate. Context-dependent interactions between proximate amino acids in a finger or binding of multiple amino acids to a single base or vice versa can cause variation of the binding specificities predicted by the existing substitution rules.

The technique of phage display provides a largely empirical means of generating
 30 zinc finger proteins with a desired target specificity (see e.g., Rebar, US 5,789,538; Choo et al., WO 96/06166; Barbas et al., WO 95/19431 and WO 98/543111; Jamieson et al.,

supra). The method can be used in conjunction with, or as an alternative to rational design. The method involves the generation of diverse libraries of mutagenized zinc finger proteins, followed by the isolation of proteins with desired DNA-binding properties using affinity selection methods. To use this method, the experimenter typically proceeds as follows. First, a gene for a zinc finger protein is mutagenized to introduce diversity into regions important for binding specificity and/or affinity. In a typical application, this is accomplished via randomization of a single finger at positions -1, +2, +3, and +6, and sometimes accessory positions such as +1, +5, +8 and +10. Next, the mutagenized gene is cloned into a phage or phagemid vector as a fusion with gene III of a filamentous phage, which encodes the coat protein pIII. The zinc finger gene is inserted between segments of gene III encoding the membrane export signal peptide and the remainder of pIII, so that the zinc finger protein is expressed as an amino-terminal fusion with pIII or in the mature, processed protein. When using phagemid vectors, the mutagenized zinc finger gene may also be fused to a truncated version of gene III encoding, minimally, the C-terminal region required for assembly of pIII into the phage particle. The resultant vector library is transformed into *E. coli* and used to produce filamentous phage which express variant zinc finger proteins on their surface as fusions with the coat protein pIII. If a phagemid vector is used, then this step requires superinfection with helper phage. The phage library is then incubated with target DNA site, and affinity selection methods are used to isolate phage which bind target with high affinity from bulk phage. Typically, the DNA target is immobilized on a solid support, which is then washed under conditions sufficient to remove all but the tightest binding phage. After washing, any phage remaining on the support are recovered via elution under conditions which disrupt zinc finger – DNA binding. Recovered phage are used to infect fresh *E. coli*, which is then amplified and used to produce a new batch of phage particles. Selection and amplification are then repeated as many times as is necessary to enrich the phage pool for tight binders such that these may be identified using sequencing and/or screening methods. Although the method is illustrated for pIII fusions, analogous principles can be used to screen ZFP variants as pVIII fusions.

In certain embodiments, the sequence bound by a particular zinc finger protein is determined by conducting binding reactions (see, *e.g.*, conditions for determination of K_d ,

infra) between the protein and a pool of randomized double-stranded oligonucleotide sequences. The binding reaction is analyzed by an electrophoretic mobility shift assay (EMSA), in which protein-DNA complexes undergo retarded migration in a gel and can be separated from unbound nucleic acid. Oligonucleotides which have bound the finger
 5 are purified from the gel and amplified, for example, by a polymerase chain reaction. The selection (*i.e.* binding reaction and EMSA analysis) is then repeated as many times as desired, with the selected oligonucleotide sequences. In this way, the binding specificity of a zinc finger protein having a particular amino acid sequence is determined.

Zinc finger proteins are often expressed with a heterologous domain as fusion
 10 proteins. Common domains for addition to the ZFP include, e.g., transcription factor domains (activators, repressors, co-activators, co-repressors), silencers, oncogenes (e.g., myc, jun, fos, myb, max, mad, rel, ets, bcl, myb, mos family members etc.); DNA repair enzymes and their associated factors and modifiers; DNA rearrangement enzymes and their associated factors and modifiers; chromatin associated proteins and their modifiers
 15 (e.g. kinases, acetylases and deacetylases); and DNA modifying enzymes (e.g., methyltransferases, topoisomerases, helicases, ligases, kinases, phosphatases, polymerases, endonucleases) and their associated factors and modifiers. A preferred domain for fusing with a ZFP when the ZFP is to be used for repressing expression of a target gene is a KRAB repression domain from the human KOX-1 protein (Thiesen et al.,
 20 *New Biologist* 2, 363-374 (1990); Margolin et al., *Proc. Natl. Acad. Sci. USA* 91, 4509-4513 (1994); Pengue et al., *Nucl. Acids Res.* 22:2908-2914 (1994); Witzgall et al., *Proc. Natl. Acad. Sci. USA* 91, 4514-4518 (1994). Preferred domains for achieving activation include the HSV VP16 activation domain (see, e.g., Hagmann et al., *J. Virol.* 71, 5952-5962 (1997)) nuclear hormone receptors (see, e.g., Torchia et al., *Curr. Opin. Cell. Biol.*
 25 10:373-383 (1998)); the p65 subunit of nuclear factor kappa B (Bitko & Barik, *J. Virol.* 72:5610-5618 (1998) and Doyle & Hunt, *Neuroreport* 8:2937-2942 (1997)); Liu et al., *Cancer Gene Ther.* 5:3-28 (1998)), or artificial chimeric functional domains such as VP64 (Seifpal et al., *EMBO J.* 11, 4961-4968 (1992)).

An important factor in the administration of polypeptide compounds, such as the
 30 ZFPs, is ensuring that the polypeptide has the ability to traverse the plasma membrane of a cell, or the membrane of an intra-cellular compartment such as the nucleus. Cellular

membranes are composed of lipid-protein bilayers that are freely permeable to small, nonionic lipophilic compounds and are inherently impermeable to polar compounds, macromolecules, and therapeutic or diagnostic agents. However, proteins and other compounds such as liposomes have been described, which have the ability to translocate polypeptides such as ZFPs across a cell membrane.

For example, “membrane translocation polypeptides” have amphiphilic or hydrophobic amino acid subsequences that have the ability to act as membrane-translocating carriers. In one embodiment, homeodomain proteins have the ability to translocate across cell membranes. The shortest internalizable peptide of a homeodomain protein, Antennapedia, was found to be the third helix of the protein, from amino acid position 43 to 58 (*see, e.g., Prochiantz, Current Opinion in Neurobiology* 6:629-634 (1996)). Another subsequence, the h (hydrophobic) domain of signal peptides, was found to have similar cell membrane translocation characteristics (*see, e.g., Lin et al., J. Biol. Chem.* 270:1 4255-14258 (1995)).

Examples of peptide sequences which can be linked to a ZFP, for facilitating uptake of ZFP into cells, include, but are not limited to: an 11 amino acid peptide of the tat protein of HIV; a 20 residue peptide sequence which corresponds to amino acids 84-103 of the p16 protein (*see Fahraeus et al., Current Biology* 6:84 (1996)); the third helix of the 60-amino acid long homeodomain of Antennapedia (Derossi *et al., J. Biol. Chem.* 269:10444 (1994)); the h region of a signal peptide such as the Kaposi fibroblast growth factor (K-FGF) h region (Lin *et al., supra*); or the VP22 translocation domain from HSV (Elliot & O’Hare, *Cell* 88:223-233 (1997)). Other suitable chemical moieties that provide enhanced cellular uptake may also be chemically linked to ZFPs.

Toxin molecules also have the ability to transport polypeptides across cell membranes. Often, such molecules are composed of at least two parts (called “binary toxins”): a translocation or binding domain or polypeptide and a separate toxin domain or polypeptide. Typically, the translocation domain or polypeptide binds to a cellular receptor, and then the toxin is transported into the cell. Several bacterial toxins, including *Clostridium perfringens* iota toxin, diphtheria toxin (DT), *Pseudomonas* exotoxin A (PE), pertussis toxin (PT), *Bacillus anthracis* toxin, and pertussis adenylate cyclase (CYA), have been used in attempts to deliver peptides to the cell cytosol as

internal or amino-terminal fusions (Arora *et al.*, *J. Biol. Chem.*, 268:3334-3341 (1993); Perelle *et al.*, *Infect. Immun.*, 61:5147-5156 (1993); Stenmark *et al.*, *J. Cell Biol.* 113:1025-1032 (1991); Donnelly *et al.*, *PNAS* 90:3530-3534 (1993); Carbonetti *et al.*, *Abstr. Annu. Meet. Am. Soc. Microbiol.* 95:295 (1995); Sebo *et al.*, *Infect. Immun.* 63:3851-3857 (1995); Klimpel *et al.*, *PNAS U.S.A.* 89:10277-10281 (1992); and Novak *et al.*, *J. Biol. Chem.* 267:17186-17193 (1992)).

Such subsequences can be used to translocate ZFPs across a cell membrane. ZFPs can be conveniently fused to or derivatized with such sequences. Typically, the translocation sequence is provided as part of a fusion protein. Optionally, a linker can be used to link the ZFP and the translocation sequence. Any suitable linker can be used, e.g., a peptide linker.

III. Position Dependence Of Subsite Recognition By Zinc Fingers

A number of the polypeptides disclosed herein have been characterized using the methods disclosed in parent application Serial No. 09/716,637 (the disclosure of which is hereby incorporated by reference in its entirety); in particular with respect to the effect of their position, within a multi-finger protein, on their sequence specificity. The results of these investigations provide a set of zinc finger sequences that are optimized for recognition of certain triplet target subsites whose 5'-most nucleotide is a G (*i.e.*, GNN triplet subsites). Thus, particular zinc finger sequences which recognize each of the GNN triplet subsites, from each position of a three-finger zinc finger protein, are provided. See Figure 2. It will be clear to those of skill in the art that the optimized, position-specific zinc finger sequences disclosed herein for recognition of GNN target subsites are not limited to use in three-finger proteins. For example, they are also useful in six-finger proteins, which can be made by linkage of two three-finger proteins.

A number of zinc finger amino acid sequences which are reported to bind to target subsites in which the 5'-most nucleotide residue is G (*i.e.*, GNN subsites) have recently been disclosed. Segal *et al.* (1999) *Proc. Natl. Acad. Sci. USA* 96:2758-2763; Drier *et al.* (2000) *J. Mol. Biol.* 303:489-502; U.S. Patent No. 6,140,081. These GNN-binding zinc fingers were obtained by selection of finger 2 sequences from phage display libraries of three-finger proteins, in which certain amino acid residues of finger 2 had been

randomized. Due to the manner in which they were selected, it is not clear whether these sequences would have the same target subsite specificity if they were present in the F1 and/or F3 positions.

Use of the methods and compositions disclosed herein has now allowed
5 identification of specific zinc finger sequences that bind each of the 16 GNN triplet subsites, and for the first time, provides zinc finger sequences that are optimized for recognition of these triplet subsites in a position-dependent fashion. Moreover, *in vivo* studies of these optimized designs reveal that the functionality of a ZFP is correlated with its binding affinity to its target sequence. See Example 6, *infra*.

10 As a result of the discovery, disclosed herein, that sequence recognition by zinc fingers is position-dependent, it is clear that existing design rules will not, in and of themselves, be applicable to every situation in which it is necessary to construct a sequence-specific ZFP. The results disclosed herein show that many zinc fingers that are constructed based on design rules exhibit the sequence specificity predicted by those
15 design rules only at certain finger positions. The position-specific zinc fingers disclosed herein are likely to function more efficiently *in vivo* and in cultured cells, with fewer nonspecific effects. Highly specific ZFPs, made using position-specific zinc fingers, will be useful tools in studying gene function and will find broad applications in areas as diverse as human therapeutics and plant engineering.

20 **IV. Production of Zinc Finger Proteins**

ZFP polypeptides and nucleic acids encoding the same can be made using routine techniques in the field of recombinant genetics. Basic texts disclosing the general methods include Sambrook et al., *Molecular Cloning, A Laboratory Manual* (2nd ed.
25 1989); Kriegler, *Gene Transfer and Expression: A Laboratory Manual* (1990); and *Current Protocols in Molecular Biology* (Ausubel et al., eds., 1994)). In addition, nucleic acids less than about 100 bases can be custom ordered from any of a variety of commercial sources, such as The Midland Certified Reagent Company (mcrc@oligos.com), The Great American Gene Company (<http://www.genco.com>),
30 ExpressGen Inc. (www.expressgen.com), Operon Technologies Inc. (Alameda, CA). Similarly, peptides can be custom ordered from any of a variety of sources, such as

PeptidoGenic (pkim@ccnet.com), HTI Bio-products, inc. (<http://www.htibio.com>), BMA Biomedicals Ltd (U.K.), Bio.Synthesis, Inc.

Oligonucleotides can be chemically synthesized according to the solid phase phosphoramidite triester method first described by Beaucage & Caruthers, *Tetrahedron Letts.* 22:1859-1862 (1981), using an automated synthesizer, as described in Van Devanter et al., *Nucleic Acids Res.* 12:6159-6168 (1984). Purification of oligonucleotides is by either denaturing polyacrylamide gel electrophoresis or by reverse phase HPLC. The sequence of the cloned genes and synthetic oligonucleotides can be verified after cloning using, e.g., the chain termination method for sequencing double-stranded templates of Wallace et al., *Gene* 16:21-26 (1981).

Two alternative methods are typically used to create the coding sequences required to express newly designed DNA-binding peptides. One protocol is a PCR-based assembly procedure that utilizes six overlapping oligonucleotides (Fig. 1). Three oligonucleotides (oligos 1, 3, and 5 in Figure 1) correspond to “universal” sequences that encode portions of the DNA-binding domain between the recognition helices. These oligonucleotides typically remain constant for all zinc finger constructs. The other three “specific” oligonucleotides (oligos 2, 4, and 6 in Fig. 1) are designed to encode the recognition helices. These oligonucleotides contain substitutions primarily at positions - 1, 2, 3 and 6 on the recognition helices making them specific for each of the different DNA-binding domains.

The PCR synthesis is carried out in two steps. First, a double stranded DNA template is created by combining the six oligonucleotides (three universal, three specific) in a four cycle PCR reaction with a low temperature annealing step, thereby annealing the oligonucleotides to form a DNA “scaffold.” The gaps in the scaffold are filled in by high-fidelity thermostable polymerase, the combination of Taq and Pfu polymerases also suffices. In the second phase of construction, the zinc finger template is amplified by external primers designed to incorporate restriction sites at either end for cloning into a shuttle vector or directly into an expression vector.

An alternative method of cloning the newly designed DNA-binding proteins relies on annealing complementary oligonucleotides encoding the specific regions of the desired ZFP. This particular application requires that the oligonucleotides be

phosphorylated prior to the final ligation step. This is usually performed before setting up the annealing reactions. In brief, the “universal” oligonucleotides encoding the constant regions of the proteins (oligos 1, 2 and 3 of above) are annealed with their complementary oligonucleotides. Additionally, the “specific” oligonucleotides encoding the finger recognition helices are annealed with their respective complementary oligonucleotides. These complementary oligos are designed to fill in the region which was previously filled in by polymerase in the above-mentioned protocol. The complementary oligos to the common oligos 1 and finger 3 are engineered to leave overhanging sequences specific for the restriction sites used in cloning into the vector of choice in the following step. The second assembly protocol differs from the initial protocol in the following aspects: the “scaffold” encoding the newly designed ZFP is composed entirely of synthetic DNA thereby eliminating the polymerase fill-in step, additionally the fragment to be cloned into the vector does not require amplification. Lastly, the design of leaving sequence-specific overhangs eliminates the need for restriction enzyme digests of the inserting fragment. Alternatively, changes to ZFP recognition helices can be created using conventional site-directed mutagenesis methods.

Both assembly methods require that the resulting fragment encoding the newly designed ZFP be ligated into a vector. Ultimately, the ZFP-encoding sequence is cloned into an expression vector. Expression vectors that are commonly utilized include, but are not limited to, a modified pMAL-c2 bacterial expression vector (New England BioLabs or an eukaryotic expression vector, pcDNA (Promega). The final constructs are verified by sequence analysis.

Any suitable method of protein purification known to those of skill in the art can be used to purify ZFPs (see, Ausubel, supra, Sambrook, supra). In addition, any suitable host can be used for expression, e.g., bacterial cells, insect cells, yeast cells, mammalian cells, and the like.

Expression of a zinc finger protein fused to a maltose binding protein (MBP-ZFP) in bacterial strain JM109 allows for straightforward purification through an amylose column (NEB). High expression levels of the zinc finger chimeric protein can be obtained by induction with IPTG since the MBP-ZFP fusion in the pMal-c2 expression plasmid is under the control of the tac promoter (NEB). Bacteria containing the MBP-

ZFP fusion plasmids are inoculated into 2xYT medium containing 10 μ M ZnCl₂, 0.02% glucose, plus 50 μ g/ml ampicillin and shaken at 37°C. At mid-exponential growth IPTG is added to 0.3 mM and the cultures are allowed to shake. After 3 hours the bacteria are harvested by centrifugation, disrupted by sonication or by passage through a french pressure cell or through the use of lysozyme, and insoluble material is removed by centrifugation. The MBP-ZFP proteins are captured on an amylose-bound resin, washed extensively with buffer containing 20 mM Tris-HCl (pH 7.5), 200 mM NaCl, 5 mM DTT and 50 μ M ZnCl₂, then eluted with maltose in essentially the same buffer (purification is based on a standard protocol from NEB). Purified proteins are quantitated and stored for biochemical analysis.

The dissociation constants of the purified proteins, e.g., K_d, are typically characterized via electrophoretic mobility shift assays (EMSA) (Buratowski & Chodosh, in *Current Protocols in Molecular Biology* pp. 12.2.1-12.2.7 (Ausubel ed., 1996)). Affinity is measured by titrating purified protein against a fixed amount of labeled double-stranded oligonucleotide target. The target typically comprises the natural binding site sequence flanked by the 3 bp found in the natural sequence and additional, constant flanking sequences. The natural binding site is typically 9 bp for a three-finger protein and 2 x 9 bp + intervening bases for a six finger ZFP. The annealed oligonucleotide targets possess a 1 base 5' overhang which allows for efficient labeling of the target with T4 phage polynucleotide kinase. For the assay the target is added at a concentration of 1 nM or lower (the actual concentration is kept at least 10-fold lower than the expected dissociation constant), purified ZFPs are added at various concentrations, and the reaction is allowed to equilibrate for at least 45 min. In addition the reaction mixture also contains 10 mM Tris (pH 7.5), 100 mM KCl, 1 mM MgCl₂, 0.1 mM ZnCl₂, 5 mM DTT, 10% glycerol, 0.02% BSA. (NB: in earlier assays poly d(IC) was also added at 10-100 μ g/ μ l.)

The equilibrated reactions are loaded onto a 10% polyacrylamide gel, which has been pre-run for 45 min in Tris/glycine buffer, then bound and unbound labeled target is resolved by electrophoresis at 150V. (alternatively, 10-20% gradient Tris-HCl gels, containing a 4% polyacrylamide stacker, can be used) The dried gels are visualized by

autoradiography or phosphorimaging and the apparent K_d is determined by calculating the protein concentration that gives half-maximal binding.

The assays can also include determining active fractions in the protein preparations. Active fractions are determined by stoichiometric gel shifts where proteins
5 are titrated against a high concentration of target DNA. Titrations are done at 100, 50, and 25% of target (usually at micromolar levels).

V. Applications of Engineered Zinc Finger Proteins

ZFPs that bind to a particular target gene, and the nucleic acids encoding them,
10 can be used for a variety of applications. These applications include therapeutic methods in which a ZFP or a nucleic acid encoding it is administered to a subject and used to modulate the expression of a target gene within the subject. *See*, for example, co-owned WO 00/41566. The modulation can be in the form of repression, for example, when the target gene resides in a pathological infecting microorganisms, or in an endogenous gene
15 of the patient, such as an oncogene or viral receptor, that is contributing to a disease state. Alternatively, the modulation can be in the form of activation when activation of expression or increased expression of an endogenous cellular gene can ameliorate a diseased state. For such applications, ZFPs, or more typically, nucleic acids encoding them are formulated with a pharmaceutically acceptable carrier as a pharmaceutical
20 composition.

Pharmaceutically acceptable carriers are determined in part by the particular composition being administered, as well as by the particular method used to administer the composition. (*see, e.g., Remington's Pharmaceutical Sciences*, 17th ed. 1985)). The ZFPs, alone or in combination with other suitable components, can be made into aerosol
25 formulations (i.e., they can be "nebulized") to be administered via inhalation. Aerosol formulations can be placed into pressurized acceptable propellants, such as dichlorodifluoromethane, propane, nitrogen, and the like. Formulations suitable for parenteral administration, such as, for example, by intravenous, intramuscular, intradermal, and subcutaneous routes, include aqueous and non-aqueous, isotonic sterile
30 injection solutions, which can contain antioxidants, buffers, bacteriostats, and solutes that render the formulation isotonic with the blood of the intended recipient, and aqueous and

non-aqueous sterile suspensions that can include suspending agents, solubilizers, thickening agents, stabilizers, and preservatives. Compositions can be administered, for example, by intravenous infusion, orally, topically, intraperitoneally, intravesically or intrathecally. The formulations of compounds can be presented in unit-dose or multi-dose sealed containers, such as ampules and vials. Injection solutions and suspensions can be prepared from sterile powders, granules, and tablets of the kind previously described.

The dose administered to a patient should be sufficient to effect a beneficial therapeutic response in the patient over time. The dose is determined by the efficacy and K_d of the particular ZFP employed, the target cell, and the condition of the patient, as well as the body weight or surface area of the patient to be treated. The size of the dose also is determined by the existence, nature, and extent of any adverse side-effects that accompany the administration of a particular compound or vector in a particular patient

In other applications, ZFPs are used in diagnostic methods for sequence specific detection of target nucleic acid in a sample. For example, ZFPs can be used to detect variant alleles associated with a disease or phenotype in patient samples. As an example, ZFPs can be used to detect the presence of particular mRNA species or cDNA in a complex mixtures of mRNAs or cDNAs. As a further example, ZFPs can be used to quantify copy number of a gene in a sample. For example, detection of loss of one copy of a p53 gene in a clinical sample is an indicator of susceptibility to cancer. In a further example, ZFPs are used to detect the presence of pathological microorganisms in clinical samples. This is achieved by using one or more ZFPs specific to genes within the microorganism to be detected. A suitable format for performing diagnostic assays employs ZFPs linked to a domain that allows immobilization of the ZFP on an ELISA plate. The immobilized ZFP is contacted with a sample suspected of containing a target nucleic acid under conditions in which binding can occur. Typically, nucleic acids in the sample are labeled (e.g., in the course of PCR amplification). Alternatively, unlabelled probes can be detected using a second labelled probe. After washing, bound-labelled nucleic acids are detected.

ZFPs also can be used for assays to determine the phenotype and function of gene expression. Current methodologies for determination of gene function rely primarily

upon either overexpression or removing (knocking out completely) the gene of interest from its natural biological setting and observing the effects. The phenotypic effects observed indicate the role of the gene in the biological system.

One advantage of ZFP-mediated regulation of a gene relative to conventional knockout analysis is that expression of the ZFP can be placed under small molecule control. By controlling expression levels of the ZFPs, one can in turn control the expression levels of a gene regulated by the ZFP to determine what degree of repression or stimulation of expression is required to achieve a given phenotypic or biochemical effect. This approach has particular value for drug development. By putting the ZFP under small molecule control, problems of embryonic lethality and developmental compensation can be avoided by switching on the ZFP repressor at a later stage in mouse development and observing the effects in the adult animal. Transgenic mice having target genes regulated by a ZFP can be produced by integration of the nucleic acid encoding the ZFP at any site *in trans* to the target gene. Accordingly, homologous recombination is not required for integration of the nucleic acid. Further, because the ZFP is trans-dominant, only one chromosomal copy is needed and therefore functional knock-out animals can be produced without backcrossing.

All references cited above are hereby incorporated by reference in their entirety for all purposes.

EXAMPLES

Example 1: Initial design of zinc finger proteins and determination of binding affinity

Initial ZFP designs were based on existing design rules, correspondence regimes and ZFP directories, including those disclosed herein (*see* Tables 1-5) and also in WO 98/53058; WO 98/530059; WO 98/53060 and co-owned US patent application Serial No. 09/444,241. *See* also WO 00/42219. Amino acid sequences were conceptually designed using amino acids 532-624 of the human transcription factor Sp1 as a backbone. Polynucleotides encoding designed ZFPs were assembled using a Polymerase Chain Reaction (PCR)-based procedure that utilizes six overlapping oligonucleotides. PCR products were directly cloned cloning into the Tac promoter

vector, pMal-c2 (New England Biolabs, Beverly, MA) using the KpnI and BamHI restriction sites. The encoded maltose binding protein-ZFP fusion polypeptides were purified according to the manufacturer's procedures (New England Biolabs, Beverly, MA). Binding affinity was measured by gel mobility-shift analysis. All of these
5 procedures are described in detail in co-owned WO 00/41566 and WO 00/42219, as well as in Zhang *et al.* (2000) *J. Biol. Chem.* **275**:33,850-33,860 and Liu *et al.* (2001) *J. Biol. Chem.* **276**:11,323-11,334; the disclosures of which are hereby incorporated by reference in their entirety.

10 **Example 2: Optimization of binding specificity by site selection**

Designed ZFPs were tested for binding specificity using site selection methods disclosed in parent application USSN 09/716,637. Briefly, designed proteins were incubated with a population of labeled, double-stranded oligonucleotides comprising a library of all possible 9- or 10-nucleotide target sequences. Five nanomoles of labeled
15 oligonucleotides were incubated with protein, at a protein concentration 4-fold above its K_d for its target sequence. The mixture was subjected to gel electrophoresis, and bound oligonucleotides were identified by mobility shift, and extracted from the gel. The purified bound oligonucleotides were amplified, and the amplification products were used for a subsequent round of selection. At each round of selection, the protein
20 concentration was decreased by 2 fold. After 3-5 rounds of selection, amplification products were cloned into the TOPO TA cloning vector (Invitrogen, Carlsbad, CA), and the nucleotide sequences of approximately 20 clones were determined. The identities of the target sites bound by a designed protein were determined from the sequences and expressed as a compilation of subsite binding sequences.

25

Example 3: Comparison of site selection results with binding affinity

To test the correlation between site selection results and the affinity of binding of a ZFP to various related targets, site selection experiments were conducted on 2 three-finger ZFPs, denoted ZFP1 and ZFP2, and the site selection results were compared with
30 K_d measurements obtained from quantitative gel-mobility shift assays using the same ZFPs and target sites. Each ZFP was constructed, based on design rules, to bind to a

particular nine-nucleotide target sequence (comprising 3 three-nucleotide subsites), as shown in Figure 1. Site selection results and affinity measurements are also shown in Figure 1. The site selection results showed that fingers 1 and 3 of both the ZFP1 and ZFP2 proteins preferentially selected their intended target sequences. However, the second finger of each ZFP preferentially selected subsites other than those to which they were designed to bind (*e.g.*, F2 of ZFP1 was designed to bind TCG, but preferentially selected GTG; F2 of ZFP2 was designed to bind GGT, but preferentially selected GGA).

To confirm the site selection results, binding affinities of ZFP1 and ZFP2 were measured (see Example 1, *supra*), both to their original target sequences and to new target sequences reflecting the site selection results. For example, the Mt-1 sequence contains two base changes (compared to the original target sequence for ZFP1) which result in a change in the sequence of the finger 2 subsite to GTG, reflecting the preferred finger 2 subsite sequence obtained by site selection. In agreement with the site selection results, binding of ZFP1 to the Mt-1 sequence is approximately 4-fold stronger than its binding to the original target sequence (K_d of 12.5 nM compared to a K_d of 50 nM, see Figure 1).

For ZFP2, the specificity of finger 2 for the 3' base of its target subsite was tested, since, although this finger was designed to bind GGT, site selection indicated that it bound preferentially to GGA. Moreover, the site selection results predicted that finger 2 of ZFP2 would bind with approximately equal affinity to GGT and GGC. Accordingly, target sequences containing GGA (Mt-3) and GGC (Mt-4) at the finger 2 subsite were constructed, and binding affinities of ZFP2 to these target sequences, and to its original target sequence (containing GGT at the finger 2 subsite), were compared. In complete agreement with the site selection results, ZFP2 exhibited the strongest binding affinity for the target sequence containing GGA at the finger 2 subsite (K_d of 0.5 nM, Figure 1), and its affinity for target sequences containing either GGT or GGC at the finger 2 subsite was approximately equal (K_d of 1 nM for both targets, Figure 1). Accordingly, the site selection method, in addition to being useful for iterative optimization of binding specificity, can also be used as a useful indicator of binding affinity.

Example 4: Use of site selection to identify position-dependent, GNN-binding zinc fingers

A large number of engineered ZFPs have been evaluated, by site selection, to identify zinc fingers that bind to GNN target subsites. In the course of these studies, it became apparent that the binding specificity of a particular zinc finger sequence is, in some instances, dependent upon the position of the zinc finger in the protein, and hence upon the location of the target subsite within the target sequence. For example, if one wishes to design a three-finger zinc finger protein to bind to a target sequence containing the triplet subsite GAT, it is necessary to know whether this subsite is the first, second or third subsite in the target sequence (*i.e.*, whether the GAT subsite will be bound by the first, second or third finger of the protein). Accordingly, over 110 three-finger zinc finger proteins, containing potential GNN-recognizing zinc fingers in various locations, have been evaluated by site selection experiments. Generally, several zinc finger sequences were designed to recognize each GNN triplet, and each design was tested in each of the F1, F2 and F3 positions through 4 to 6 rounds of selection.

The results of these analyses, shown in Figure 2, provide optimal position-dependent zinc finger sequences (the sequences shown represent amino acid residues –1 through +6 of the recognition helix portion of the finger) for recognition of the 16 GNN target subsites, as well as site selection results for these GNN-specific zinc fingers.

Optimal amino acid sequences for recognition of each GNN subsite from each of three positions (finger 1, finger 2 or finger 3) are thereby provided.

GNG-binding finger designs

The amino acid sequence RSDXLXR (position –1 to +6 of the recognition helix) was found to be optimal for binding to the four GNG triplets, with Asn⁺³ specifying A as the middle nucleotide; His⁺³ specifying G as the middle nucleotide; Ala⁺³ specifying T as the middle nucleotide; and Asp⁺³ specifying cytosine as the middle nucleotide. At the +5 position, Ala, Thr, Ser, and Gln, were tested, and all showed similar specificity profiles by site selection. Interestingly, and in contrast to a previous report (Swirnoff *et al.* (1995) *Mol. Cell. Biol.* **15**:2275-2287), site selection results indicated that three naturally-occurring GCG-binding fingers from zif268 and Sp1, having the amino acid sequences RSEDLTR, RSEDLQR, and RSDERKR, were not GCG-specific. Rather, each of these

fingers selected almost equal numbers of GCG and GTG sequences. Analysis of binding affinity by gel-shift experiments confirmed that finger 3 of zif268, having the sequence RSDERKR, binds GCG and GTG with approximately equal affinity.

Position dependence of GCA-, GAT-, GGT-, GAA- and GCC-binding fingers

Based on existing design rules, the amino acid sequence QSGDLTR (-1 through +6) was tested for its ability to bind the GCA triplet from three positions (F1, F2, and F3) within a three-finger ZFP. Figure 3A shows that the QSGDLTR sequence bound preferentially to the GCA triplet subsite from the F2 and F3 positions, but not from F1. In fact, the presence of QSGDLTR at the F1 position of three different three-finger ZFPs resulted predominantly in selection of GCT. Accordingly, an attempt was made to redesign this sequence to obtain specificity for GCA from the F1 position. Since the sequence $Q^{-1}G^{+2}S^{+3}R^{+6}$ had previously been selected from a randomized F1 library using GCA as target (Rebar *et al.* (1994) *Science* **263**:671-673), a D (asp) to S (ser) change was made at the +3 residue of this finger. The resulting sequence, QSGSLTR, was tested for its binding specificity by site selection and found to preferentially bind GCA, from the F1 position, in three different ZFPs (see Figure 2).

The QSGSLTR zinc finger, optimized for recognition of the GCA subsite from the F1 position, was tested for its selectivity when located at the F2 position. Accordingly, two ZFPs, one containing QSGSLTR at finger 2 and one containing QSGDLTR at finger 2 (both having identical F1 sequences and identical F3 sequences) were tested by site selection. The results indicated that, when used at the F2 position, QSGSLTR bound preferentially to GTA, rather than GCA. Thus, for optimal binding of a GCA triplet subsite from the F1 position, the amino acid sequence QSGSLTR is required; while, for optimal binding of the same subsite sequence from F2 or F3, QSGDLTR should be used. Accordingly, different zinc finger amino acid sequences may be needed to specify a particular triplet subsite sequence, depending upon the location of the subsite within the target sequence and, hence, upon the position of the finger in the protein.

Positional effects were also observed for zinc fingers recognizing GAT and GGT subsites. The zinc finger amino acid sequence QSSNLAR (-1 through +6) is expected to bind to GAT, based on design rules. However, this sequence selected GAT only from the

F1 position, and not from the F2 and F3 positions, from which the sequence GAA was preferentially bound (Figure 3B). Similarly, the amino acid sequence QSSHLTR which, based on design rules, should bind GGT, selected GGT at the F1 position, but not at the F2 and F3 positions, from which it preferentially bound GGA (Figure 3C). Conversely, the amino acid sequence TSGHLVR has previously been disclosed to recognize the triplet GGT, based on its selection from a randomized library of zif268 finger 2. U.S. Patent No. 6,140,081. However, TSGHLVR was not specific for the GGT subsite when located at the F1 position (Figure 3C). These results indicate that the binding specificity of many fingers is position dependent, and particularly point out that the sequence specificity of a zinc finger selected from a F2 library may be positionally limited.

The results shown in Figure 2 indicate that recognition of at least GAA and GCC triplets by zinc fingers is also position dependent.

These positional dependences stand in contrast to earlier published work, which suggested that zinc fingers behaved as independent modules with respect to the sequence specificity of their binding to DNA. Desjarlais *et al.* (1993) *Proc. Natl. Acad. Sci. USA* 90:2256-2260.

Example 5: Characterization of EP2C

The engineered zinc finger protein EP2C binds to a target sequence, GCGGTGGCT with a dissociation constant (K_d) of 2 nM. Site selection results indicated that fingers 1 and 2 are highly specific for their target subsites, while finger 3 selects GCG (its intended target subsite) and GTG at approximately equal frequencies (Figure 4A). To confirm these observations, the binding affinities of EP2C to its cognate target sequence, and to variant target sequences, was measured by standard gel-shift analyses (see Example 1, *supra*). As standards for comparison, the binding affinities of Sp1 and zif268 to their respective targets were also measured under the same conditions, and were determined to be 40 nM for SP1 (target sequence GGGGCGGGG) and 2 nM for zif268 (target sequence GCGTGGGCG). Measurements of binding affinities confirmed that F3 of EP2C bound GTG and GCG equally well (K_d s of 2 nM), but bound GAG with a two-fold lower affinity (Figure 4B). Finger 2 was very specific for the GTG triplet, binding 15-fold less tightly to a GGG triplet (compare 2C0 and 2C3 in Figure 4B).

Finger 1 was also very specific for the GCT triplet, it bound with 4-fold lower affinity to a GAT triplet (2C4) and with 2-fold lower affinity to a GCG triplet (2C5). This example shows, once again, the high degree of correlation between site selection results and binding affinities.

5

Example 6: Evaluation of engineered ZFPs by *in vivo* functional assays

To determine whether a correlation exists between the binding affinity of an engineered ZFP to its target sequence and its functionality *in vivo*, cell-based reporter gene assays were used to analyze the functional properties of the engineered ZFP EP2C (see Example 5, *supra*). For these assays, a plasmid encoding the EP2C ZFP, fused to a VP16 transcriptional activation domain, was used to construct a stable cell line (T-Rex-293TM, Invitrogen, Carlsbad, CA) in which expression of EP2C-VP16 is inducible, as described in Zhang *et al.*, *supra*. To generate reporter constructs, three tandem copies of the EP2C target site, or its variants (see Figure 4B, column 2), were inserted between the Mlu I and BglII sites of the pGL3 luciferase-encoding vector (Promega, Madison, WI), upstream of the SV40 promoter. Structures of all reporter constructs were confirmed by DNA sequencing.

Luciferase reporter assays were performed by co-transfection of luciferase reporter construct (200 ng) and pCMV- β gal (100 ng, used as an internal control) into the EP2C cells seeded in 6-well plates. Expression of the EP2C-VP16 transcriptional activator was induced with doxycycline (0.05 μ g/ml) 24 h after transfection of reporter constructs. Cell lysates were harvested 40 hours post-transfection, luciferase and β -galactosidase activities were measured by the Dual-Light Reporter Assay System (Tropix, Bedford, MA), and luciferase activities were normalized to the co-transfected β -galactosidase activities. The results, shown on the right side of Figure 4B, showed that the normalized luciferase activity for each reporter construct was well correlated with the *in vitro* binding affinity of EP2C to the target sequence present in the construct. For example, the target sequences to which EP2C bound with greatest affinity (2C0 and 2C2, K_d of 2 nM for each) both stimulated the highest levels of luciferase activity, when used to drive luciferase expression in the reporter construct (Figure 4B). Target sequences to which EP2C bound with 2-fold lower affinity, 2C1 and 2C5 (K_d of 4 nM for each),

stimulated roughly half the luciferase activity of the 2C0 and 2C2 targets. The 2C3 and 2C4 sequences, for which EP2C showed the lowest *in vitro* binding affinities, also yielded the lowest levels of *in vivo* activity when used to drive luciferase expression. Target 3B, a sequence to which EP2C does not bind, yielded background levels of

5 luciferase activity, similar to those obtained with a luciferase-encoding vector lacking EP2C target sequences (pGL3). Thus there exist good correlations between binding affinity (as determined by K_d measurement), binding specificity (as determined by site selection) and *in vivo* functionality for engineered zinc finger proteins.

10

FOOT " 4558880

TABLE 1

<u>SBS#</u>	<u>TARGET</u>	SEQ ID	<u>F1</u>	SEQ ID	<u>F2</u>	SEQ ID	<u>F3</u>	SEQ ID	<u>Kd</u> (nM)
249	GCGGGGGCG	17	RSDELTR	123	RSDHLSR	229	RSDELRR	335	20
250	GCGGGGGCG	18	RSDELTR	124	RSDHLSR	230	RSDTLKK	336	70
251	GCGGAGGCG	19	RSDELTR	125	RSDNLTR	231	RSDELRR	337	27.5
252	GCGGCCGCG	20	RSDELTR	126	DRSSLTR	232	RSDELRR	338	100
253	GGATGGGGG	21	RSDHLAR	127	RSDHLTT	233	QRAHLAR	339	0.75
256	GCGGGGTCC	22	ERGDLTT	128	RSDHLSR	234	RSDELRR	340	800
258	GCGGGCGGG	23	RSDHLTR	129	ERGHLTR	235	RSDELRR	341	15
259	GCAGAGGAG	24	RSDNLAR	130	RSDNLAR	236	QSGSLTR	342	250
261	GAGGTGGCC	25	ERGTLAR	131	RSDALSR	237	RSDNLSR	343	0.5
262	GCGGGGGCT	26	QSSDLQR	132	RSDHLSR	238	RSDELRR	344	20
263	GCGGGGGCT	27	QSSDLQR	133	RSDHLSR	239	RSDTLKK	345	1
264	GTGGCTGCC	28	DRSSLTR	134	QSSDLQR	240	RSDALAR	346	27
265	GTGGCTGCC	29	ERGTLAR	135	QSSDLQR	241	RSDALAR	347	600
269	GGGGCCGGG	30	RSDHLTR	136	DRSSLTR	242	RSDHLTR	348	5
270	GGGGCCGGG	31	RSDHLTR	137	ERGTLAR	243	RSDHLTR	349	52.5
272	GCAGGGGCC	32	DRSSLTR	138	RSDHLSR	244	QSGSLTR	350	20
337	TGCGGGGCAA	33	RSADLTR	139	RSDHLTR	245	ERQHLAT	351	24
338	TGCGGGGCAA	34	RSADLTR	140	RSDHLTR	246	ERDHLRT	352	8
339	TGCGGGGCAA	35	RSADLTR	141	RSDHLTT	247	ERQHLAT	353	64
340	TGCGGGGCAA	36	RSADLTR	142	RSDHLTT	248	ERDHLRT	354	48
341	TGCGGGGCAA	37	RSADLTR	143	RGDHLKD	249	ERQHLAT	355	1000
342	TGCGGGGCAA	38	RSADLTR	144	RGDHLKD	250	ERDHLRT	356	1000
343	TGCGGGGCAA	39	QSGSLTR	145	RSDHLTR	251	ERQHLAT	357	8
344	TGCGGGGCAA	40	QSGSLTR	146	RSDHLTR	252	ERDHLRT	358	6

"T002T" 466666

345	TGCGGGGCAA	41	QSGSLTR	147	RSDHLTT	253	ERQHLAT	359	96
346	TGCGGGGCAA	42	QSGSLTR	148	RSDHLTT	254	ERDHLRT	360	64
347	TGCGGGGCAA	43	QSGSLTR	149	RGDHLKD	255	ERQHLAT	361	1000
348	TGCGGGGCAA	44	QSGSLTR	150	RGDHLKD	256	ERDHLRT	362	1000
367	GGGGGCGGG	45	RSDHLTR	151	DSGHLTR	257	RSDHLQR	363	60
368	GAGGGGGCG	46	RSDELTR	152	RSDHLTR	258	RSDNLTR	364	3.5
369	GTAGTTGTG	47	RSDALTR	153	TGGSLAR	259	QSGSLTR	365	95
370	GTAGTTGTG	48	RSDALTR	154	NRATLAR	260	QSASLTR	366	300
371	GTAGTTGTG	49	RSDALTR	155	NRATLAR	261	QSGSLTR	367	175
372	GTAGTTGTG	50	RSDSLLR	156	TGGSLAR	262	QSASLTR	368	112.5
373	GTAGTTGTG	51	RSDSLLR	157	NRATLAR	263	QSASLTR	369	320
374	GCTGAGGAA	52	QRSNLVR	158	RSDNLTR	264	TSSELQR	370	3.3
375	GAGGAAGAT	53	QQSNLAR	159	QSGNLQR	265	RSDNLTR	371	85
401	GTAGTTGTG	54	RSDALTR	160	TGGSLAR	266	QSASLTR	372	80
403	GTAGTTGTG	55	RSDSLLR	161	NRATLAR	267	QSGSLTR	373	750
421	GTAGTTGTG	56	DSDSLLR	162	TGGSLAR	268	QSGSLTR	374	500
422	GTAGTTGTG	57	RSDSLLR	163	TGGSLTR	269	QSGSLTR	375	200
423	GTAGTTGTG	58	RSDALTR	164	TGGSLAR	270	QRSALAR	376	1000
424	GATGCTGAG	59	RSDNLTR	165	TSSELQR	271	TSANLSR	377	100
425	GATGCTGAG	60	RSDNLTR	166	QSSDLQR	272	QQSNLAR	378	25
426	GATGCTGAG	61	RSDNLTR	167	QSSDLQR	273	TSANLSR	379	5.5
427	GCTGAGGAA	62	QRSNLVR	168	RSDNLTR	274	QSSDLQR	380	1
428	GAAGATGAC	63	DSSNLTR	169	QQSNLAR	275	QRSNLVR	381	120
429	GAAGATGAC	64	DSSNLTR	170	TSANLSR	276	QRSNLVR	382	50
430	GATGACGAC	65	EKANLTR	171	DSSNLTR	277	QQSNLAR	383	250
431	GACGACGGC	66	DSGHLTR	172	DRSNLER	278	DSSNLTR	384	100
432	GACGACGGC	67	DSGHLTR	173	DHANLAR	279	DSSNLTR	385	1000
433	GACGACGGC	68	DSGNLTR	174	DHANLAR	280	DSSNLTR	386	1000
434	GACGGCGTA	69	QSASLTR	175	DSGHLTR	281	EKANLTR	387	152.5
435	GACGGCGTA	70	QSASLTR	176	DSGHLTR	282	ERGNLTR	388	150
436	GACGGCGTA	71	QRSALAR	177	DSGHLTR	283	EKANLTR	389	95

0999994 1000
T002T "4666660

[illegible]

544	GAGTGTGTGA	103	RSDSLLR	209	MSHHLKE	315	RSDNLAR	421	125
545	GAGTGTGTGA	104	RKDSLVR	210	TSDHLAS	316	RSDNLTR	422	32
546	GAGTGTGTGA	105	RSDLLQR	211	MSHHLKT	317	RLDGLRT	423	500
547	GAGTGTGTGA	106	RKDSLVR	212	TSGLHTS	318	RSDNLTR	424	500
548	GAGTGTGTGA	107	RSSLLQR	213	MSHHLKT	319	RSDHLSR	425	500
549	GAGTGTGTGA	108	RSSLLQR	214	MSHHLKE	320	RSDHLSR	426	500
550	GAGTGTGTGA	109	RKDSLVR	215	TKDHLAS	321	RSDNLTR	427	20
551	GAGTGTGTGA	110	RSDLLQR	216	MSHHLKT	322	RSDHLSR	428	50
552	GAGTGTGTGA	111	RKDSLVR	217	MSHHLKT	323	RSDNLTR	429	31
553	GAGTGTGTGA	112	RSDSLLR	218	MSHHLKE	324	RSDNLTR	430	125
554	GAGTGTGTGA	113	RKDSLVR	219	TSDHLAS	325	RSDNLAR	431	62.5
558	TGCGGGGCA	114	QSGDLTR	220	RSDHLTR	326	DSGHLAS	432	21
559	GAGTGTGTGA	115	RSDSLLR	221	TSDHLAS	327	RSDNLAR	433	1000
560	GAGTGTGTGA	116	RSSLLQR	222	MSHHLKT	328	RSDHLSR	434	500
561	GAGTGTGTGA	117	RKDSLVR	223	MSHHLKE	329	RSDNLAR	435	1000
562	GAGTGTGTGA	118	RSDSLLR	224	TSGLHTS	330	RSDNLAR	436	1000
565	GATGCTGAG	119	RSDNLTR	225	TSSELQR	331	QQSNLAR	437	100
567	GAAGATGAC	120	EKANLTR	226	TSANLSR	332	QRSNLVR	438	47.5
568	GATGACGAC	121	EKANLTR	227	DSSNLTR	333	TSANLSR	439	300
569	GTAGTTGTG	122	RSDSLLR	228	TGGSLAR	334	QRSALTR	440	52

TABLE "4666660"

TABLE 2

		<u>SEQ</u>		<u>SEQ</u>		<u>SEQ</u>		<u>SEQ</u>	<u>Kd</u>
<u>SBS#</u>	<u>TARGET</u>	<u>ID</u>	<u>F1</u>	<u>ID</u>	<u>F2</u>	<u>ID</u>	<u>F3</u>	<u>ID</u>	<u>(nM)</u>
201	GCAGCCTTG	441	RSDSLTS	646	ERSTLTR	851	QRADLRR	1056	1000
202	GCAGCCTTG	442	RSDSLTS	647	ERSTLTR	852	QRADLAR	1057	1000
203	GCAGCCTTG	443	RSDSLTS	648	ERSTLTR	853	QRATLRR	1058	1000
204	GCAGCCTTG	444	RSDSLTS	649	ERSTLTR	854	QRATLAR	1059	1000
205	GAGGTAGAA	445	QSANLAR	650	QSATLAR	855	RSDNLSR	1060	80
206	GAGGTAGAA	446	QSANLAR	651	QSAVLAR	856	RSDNLSR	1061	1000
207	GAGTGGTTA	447	QRASLAS	652	RSDHLTT	857	RSDNLAR	1062	70
208	TAGGTCTTA	448	QRASLAS	653	DRSALAR	858	RSDNLAS	1063	1000
209	GGAGTGGTT	449	QSSALAR	654	RSDALAR	859	QRAHLAR	1064	35
210	GGAGTGGTT	450	NRDTLAR	655	RSDALAR	860	QRAHLAR	1065	65
211	GGAGTGGTT	451	QSSALAR	656	RSDALAS	861	QRAHLAR	1066	140
212	GGAGTGGTT	452	NRDTLAR	657	RSDALAS	862	QRAHLAR	1067	400
213	GTTGCTGGA	453	QRAHLAR	658	QSSTLAR	863	QSSALAR	1068	1000
214	GTTGCTGGA	454	QRAHLAR	659	QSSTLAR	864	NRDTLAR	1069	1000
215	GAAGTCTGT	455	NRDHLMV	660	DRSALAR	865	QSANLSR	1070	1000
216	GAAGTCTGT	456	NRDHLTT	661	DRSALAR	866	QSANLSR	1071	1000
217	GAGGTCGTA	457	QRSALAR	662	DRSALAR	867	RSDNLAR	1072	40
219	GATGTTGAT	458	QQSNLAR	663	NRDTLAR	868	NRDNLSR	1073	1000
220	GATGTTGAT	459	QQSNLAR	664	NRDTLAR	869	QQSNLSR	1074	1000
221	GATGAGTAC	460	DRSNLRT	665	RSDNLAR	870	NRDNLAR	1075	1000
222	GATGAGTAC	461	ERSNLRT	666	RSDNLAR	871	NRDNLAR	1076	1000
223	GATGAGTAC	462	DRSNLRT	667	RSDNLAR	872	QQSNLAR	1077	105
224	GATGAGTAC	463	ERSNLRT	668	RSDNLAR	873	QQSNLAR	1078	1000
225	TGGGAGGTC	464	DRSALAR	669	RSDNLAR	874	RSDHLTT	1079	6
226	GCAGCCTTG	465	RGDALTS	670	ERGTLAR	875	QSGSLTR	1080	1000
227	GCAGCCTTG	466	RGDALTV	671	ERGTLAR	876	QSGSLTR	1081	1000

T002T"466660

42

Parameter	Unit	Value	Standard Error	95% CI	P-value
Intercept		1.00	0.00	1.00	<0.001
Age	Year	0.02	0.01	-0.01, 0.05	0.15
Gender					
Male		0.01	0.02	-0.03, 0.05	0.78
Female		-0.01	0.02	-0.05, 0.03	0.82
Education	Year	0.01	0.01	-0.01, 0.03	0.45
Income	Year	0.01	0.01	-0.01, 0.03	0.45
Health status					
Good		0.01	0.02	-0.03, 0.05	0.78
Poor		-0.01	0.02	-0.05, 0.03	0.82
Smoking status					
Smoker		0.01	0.02	-0.03, 0.05	0.78
Nonsmoker		-0.01	0.02	-0.05, 0.03	0.82
Alcohol consumption					
Drinker		0.01	0.02	-0.03, 0.05	0.78
Nondrinker		-0.01	0.02	-0.05, 0.03	0.82
Family size					
Small		0.01	0.02	-0.03, 0.05	0.78
Large		-0.01	0.02	-0.05, 0.03	0.82
Marital status					
Married		0.01	0.02	-0.03, 0.05	0.78
Single		-0.01	0.02	-0.05, 0.03	0.82
Occupation					
Professional		0.01	0.02	-0.03, 0.05	0.78
Service		-0.01	0.02	-0.05, 0.03	0.82
Unemployed		0.01	0.02	-0.03, 0.05	0.78
Retired		-0.01	0.02	-0.05, 0.03	0.82
Region					
North		0.01	0.02	-0.03, 0.05	0.78
South		-0.01	0.02	-0.05, 0.03	0.82
East		0.01	0.02	-0.03, 0.05	0.78
West		-0.01	0.02	-0.05, 0.03	0.82
Season					
Spring		0.01	0.02	-0.03, 0.05	0.78
Summer		-0.01	0.02	-0.05, 0.03	0.82
Autumn		0.01	0.02	-0.03, 0.05	0.78
Winter		-0.01	0.02	-0.05, 0.03	0.82

[illegible]

TABLE 3

<u>SBS#</u>	<u>TARGET</u>	<u>SEQ</u> <u>ID</u>	<u>F1</u>	<u>SEQ</u> <u>ID</u>	<u>F2</u>	<u>SEQ</u> <u>ID</u>	<u>F3</u>	<u>SEQ</u> <u>ID</u>	<u>Kd</u> <u>(nM)</u>
897	GAGGAGGTGA	1261	RSDALAR	1347	RSDNLAR	1433	RSDNLVR	1519	0.07
828	GCGGAGGACC	1262	EKANLTR	1348	RSDNLAR	1434	RSDEKRR	1520	0.1
884	GAGGAGGTGA	1263	RSDSLTR	1349	RSDNLAR	1435	RSDNLVR	1521	0.15
817	GAGGAGGTGA	1264	RSDSLTR	1350	RSDNLAR	1436	RSDNLAR	1522	0.31
666	GCGGAGGCGC	1265	RSDDLTR	1351	RSDNLTR	1437	RSDTLKK	1523	0.5
829	GCGGAGGACC	1266	EKANLTR	1352	RSDNLAR	1438	RSDTLKK	1524	0.52
670	GACGTGGAGG	1267	RSDNLAR	1353	RSDALAR	1439	DRSNLTR	1525	0.57
801	AAGGAGTCGC	1268	RSADLRT	1354	RSDNLAR	1440	RSDNLTQ	1526	0.85
668	GTGGAGGCCA	1269	ERGTLAR	1355	RSDNLAR	1441	RSDALAR	1527	1.13
895	ATGGATTGAG	1270	QSHDLTK	1356	TSGNLVR	1442	RSDALTQ	1528	1.4
799	GGGGGAGCTG	1271	QSSDLQR	1357	QRAHLER	1443	RSDHLNR	1529	1.85
798	GGGGGAGCTG	1272	QSSDLQR	1358	QSGHLQR	1444	RSDHLNR	1530	3
842	GAGGTGGGCT	1273	DRSHLTR	1359	RSDALAR	1445	RSDNLAR	1531	5.4
894	TCAGTGGTAT	1274	QRSALAR	1360	RSDALNR	1446	QSHDLTK	1532	6.15
892	ATGGATTGAG	1275	QSHDLTK	1361	QSNLVR	1447	RSDALTQ	1533	6.2
888	TCAGTGGTAT	1276	QSSSLVR	1362	RSDALNR	1448	QSHDLTK	1534	14
739	GCGGGCGGGC	1277	RSDHLTR	1363	ERHGLTR	1449	RSDDLNR	1535	16.5
850	CAGGCTGTGG	1278	RSDALTR	1364	QSSDLTR	1450	RSDNLRE	1536	17
797	GCAGAGGCTG	1279	QSSDLQR	1365	RSDNLAR	1451	QSGDLTR	1537	17.5
891	TCAGTGGTAT	1280	QSSSLVR	1366	RSDALNR	1452	QSGSLRT	1538	18.5
887	TCAGTGGTAT	1281	QRSALAR	1367	RSDALNR	1453	QSGDLRT	1539	23.75
672	TCGGACGTGG	1282	RSDALAR	1368	DRSNLTR	1454	RSDELRT	1540	24
836	GGGGAGGCCC	1283	ERGTLAR	1369	RSDNLAR	1455	RSDHLNR	1541	24.25
674	GCGGCGTCGG	1284	RSDELRT	1370	RADTLNR	1456	RSDTLKK	1542	27.5
849	GGGGCCCTGG	1285	RSDALRE	1371	DRSSLTR	1457	RSDHLTQ	1543	29.05
825	GAATGGGCAG	1286	QSGSLTR	1372	RSDHLTT	1458	QSGNLTR	1544	37.3

T002T "16668660

673	GCGGGTGTCT	1287	DRSALAR	1373	QSSHLAR	1459	RSDTLKK	1545	48.33
848	GGGGAGGCCC	1288	DRSSLTR	1374	RSDNLAR	1460	RSDHLSR	1546	49.5
662	AGAGCGGCAC	1289	QTGSLTR	1375	RSDELQR	1461	QSGHLNQ	1547	50
667	GAGTCGGACG	1290	DRSNLTR	1376	RSDELRT	1462	RSDNLAR	1548	50
803	GCAGCGGCTC	1291	QSSDLQR	1377	RSDELQR	1463	QSGSLTR	1549	57.5
671	TCGGACGAGT	1292	RSDNLAR	1378	DRSNLTR	1464	RSDELRT	1550	64
851	GAGATGGATC	1293	QSSNLQR	1379	RRDVL MN	1465	RLHNLQR	1551	74
804	GCAGCGGCTC	1294	QSSDLQR	1380	RSDDLNR	1466	QSGSLTR	1552	82.5
669	GACGAGTCGG	1295	RSDELRT	1381	RSDNLAR	1467	DRSNLTR	1553	90
682	GCTGCAGGAG	1296	RSDHLAR	1382	QSGDLTR	1468	QSSDLSR	1554	90
845	GAGATGGATC	1297	QSSNLQR	1383	RSDALRQ	1469	RLHNLQR	1555	112.5
663	AGAGCGGCAC	1298	QTGSLTR	1384	RSDELQR	1470	KNWKLQA	1556	115
738	GCGGGGTCCG	1299	ERGTLTT	1385	RSDHLSR	1471	RSDDLRR	1557	120
664	AGAGCGGCAC	1300	QTGSLTR	1386	RADTLRR	1472	ASSRLAT	1558	125
833	GACTAGGACC	1301	EKANLTR	1387	RSDNLTK	1473	DRSNLTR	1559	136
685	GCTGCAGGAG	1302	RSDHLAR	1388	QSGSLTR	1474	QSSDLSR	1560	150
835	TAGGGAGCGT	1303	RADTLRR	1389	QSGHLTR	1475	RSDNLTT	1561	150
847	TAGGGAGCGT	1304	RSDDLTR	1390	QSGHLTR	1476	RSDNLTT	1562	150
818	GAATGGGCAG	1305	QSGSLTR	1391	RSDHLTT	1477	QSSNLVR	1563	167
834	GACTAGGACC	1306	EKANLTR	1392	RSDHLTT	1478	DRSNLTR	1564	186
837	GGGGCCCTGG	1307	RSDALRE	1393	DRSSLTR	1479	RSDHLSR	1565	222
764	GCAGAGGCTG	1308	TSGELVR	1394	RSDNLAR	1480	QSGDLTR	1566	255
774	GCAGCGGTAG	1309	QRSALAR	1395	RSDELQR	1481	QSGDLTR	1567	258
765	GCCGAGGCCG	1310	ERGTLAR	1396	RSDNLAR	1482	ERGTLAR	1568	262.5
766	GCCGAGGCCG	1311	ERGTLAR	1397	RSDNLAR	1483	DRSDLTR	1569	262.5
775	GCAGCGGTAG	1312	QSGALTR	1398	RSDELQR	1484	QSGDLTR	1570	265
763	GCAGAGGCTG	1313	TSGELVR	1399	RSDNLAR	1485	QSGSLTR	1571	275
838	GGGGCCCTGG	1314	RSDALRE	1400	DRSSLTR	1486	RSDHLTA	1572	300
841	GAGTGTGAGG	1315	RSDNLAR	1401	QSSHLAS	1487	RSDNLAR	1573	300
770	TTGGCAGCCT	1316	DRSSLTR	1402	QSGSLTR	1488	RSDSLTK	1574	325
767	GGGGGAGCTG	1317	QSSDLAR	1403	QSGHLQR	1489	RSDHLSR	1575	335

T.D. 2021-11-16 09:00

800	TTGGCAGCCT 1318	ERGTLAR 1404	QSGSLTR 1490	RSDSLTK 1576	400
832	GACTAGGACC 1319	EKANLTR 1405	RSDNLTT 1491	DRSNLTR 1577	408
844	GAGATGGATC 1320	QSSNLQR 1406	RSDALRQ 1492	RSDNLQR 1578	444
683	GCTGCAGGAG 1321	QSGHLAR 1407	QSGSLTR 1493	QSSDLSR 1579	500
805	GCAGCGGTAG 1322	QRSALAR 1408	RSDELQR 1494	QSGSLTR 1580	500
839	GAGTGTGAGG 1323	RSDNLAR 1409	TSDHLAS 1495	RSDNLAR 1581	625
840	GAGTGTGAGG 1324	RSDNLAR 1410	MSHHLKT 1496	RSDNLAR 1582	625
830	GGAGAGTCGG 1325	RSDELRT 1411	RSDNLAR 1497	QRAHLAR 1583	683
831	GGAGAGTCGG 1326	RSDDLTK 1412	RSDNLAR 1498	QRAHLAR 1584	700
684	GCTGCAGGAG 1327	RSAHLAR 1413	QSGSLTR 1499	QSSDLSR 1585	850
846	GAGATGGATC 1328	QSSNLQR 1414	RRDVLNM 1500	RSDNLQR 1586	889.5
819	AAGTAGGGTG 1329	QSSHLTR 1415	RSDNLTT 1501	RSDNLTQ 1587	1000
820	ACGGTAGTTA 1330	QSSALTR 1416	QRSALAR 1502	RSDTLTQ 1588	1000
821	ACGGTAGTTA 1331	NRATLAR 1417	QRSALAR 1503	RSDTLTQ 1589	1000
822	GTGTGCTGGT 1332	RSDHLTT 1418	ERQHLAT 1504	RSDALAR 1590	1000
823	GTGTGCTGGT 1333	RSDHLTK 1419	ERQHLAT 1505	RSDALAR 1591	1000
824	GTGTGCTGGT 1334	RSDHLTT 1420	DRSHLRT 1506	RSDALAR 1592	1000
885	GTGTGCTGGT 1335	RSDHLTK 1421	DRSHLRT 1507	RSDALAR 1593	1000
886	TCAGTGGTAT 1336	QSSSLVR 1422	RSDALSR 1508	QSGDLRT 1594	1000
889	ATGGATTGAT 1337	QSGSLTT 1423	QQSNLVR 1509	RSDALTQ 1595	1000
890	CTGGTATGTC 1338	QRSHLTT 1424	QRSALAR 1510	RSDALRE 1596	1000
896	AAGTAGGGTG 1339	TSGHLVR 1425	RSDNLTT 1511	RSDNLTQ 1597	1000
898	ACGGTAGTTA 1340	NRATLAR 1426	QSSSLVR 1512	RSDTLTQ 1598	1000
899	CTGGTATGTC 1341	QRSHLTT 1427	QSSSLVR 1513	RSDALRE 1599	1000
900	CTGGTATGTC 1342	MSHHLKE 1428	QSSSLVR 1514	RSDALRE 1600	1000
901	CTGGTATGTC 1343	MSHHLKE 1429	QRSALAR 1515	RSDALRE 1601	1000
773	GCAGCGGTAG 1344	QSGALTR 1430	RSDELQR 1516	QSGSLTR 1602	1250
768	GGGGGAGCTG 1345	QSSDLAR 1431	QRAHLER 1517	RSDHLR 1603	2000
681	GCTGCAGGAG 1346	RSAHLAR 1432	QSGDLTR 1518	QSSDLSR 1604	3000

TABLE 4

SBS#	TARGET	<u>SEQ</u> <u>ID</u>	F1	<u>SEQ</u> <u>ID</u>	F2	<u>SEQ</u> <u>ID</u>	F3	<u>SEQ</u> <u>ID</u>	<u>Kd</u> <u>(nM)</u>
607	AAGGTGGCAG	1605	QSGDLTR	1707	RSDSLAR	1809	RLDNRTA	1911	6.5
608	TTGGCTGGGC	1606	GSWHLTR	1708	QSSDLQR	1810	RSDSLTK	1912	8
611	GTGGCTGCAG	1607	QSGDLTR	1709	QSSDLQR	1811	RSDALAR	1913	11.5
612	GTGGCTGCAG	1608	QSGTLTR	1710	QSSDLQR	1812	RSDALAR	1914	0.38
613	TTGGCTGGGC	1609	RSDHLAR	1711	QSSDLQR	1813	RGDALTS	1915	1.45
614	TTGGCTGGGC	1610	RSDHLAR	1712	QSSDLQR	1814	RSDSLTK	1916	2
616	GAGGAGGATG	1611	QSSNLQR	1713	RSDNLAR	1815	RSDNLQR	1917	0.08
617	AAGGGGGGG	1612	RSDHLSR	1714	RSDHLTR	1816	RKDNMTA	1918	1
618	AAGGGGGGG	1613	RSDHLSR	1715	RSDHLTR	1817	RKDNMTQ	1919	0.55
619	AAGGGGGGG	1614	RSDHLSR	1716	RSDHLTR	1818	RKDNMTN	1920	1.34
620	AAGGGGGGG	1615	RSDHLSR	1717	RSDHLTR	1819	RLDNRTA	1921	0.54
621	AAGGGGGGG	1616	RSDHLSR	1718	RSDHLTR	1820	RLDNRTQ	1922	0.75
624	ACGGATGTCT	1617	DRSALAR	1719	TSANLAR	1821	RSDTLRS	1923	7
628	TTGTAGGGGA	1618	RSDHLTR	1720	RSDNLTT	1822	RGDALTS	1924	130
629	TTGTAGGGGA	1619	RSSHLTR	1721	RSDNLTT	1823	RGDALTS	1925	150
630	CGGGGAGAGT	1620	RSDNLAR	1722	QSGHLQR	1824	RSDHLRE	1926	37.5
646	TTGGTGGAAG	1621	QSGNLAR	1723	RSDALAR	1825	RGDALTS	1927	35
647	TTGGTGGAAG	1622	QSANLAR	1724	RSDALAR	1826	RGDALTS	1928	40
651	GTTGTGGAAT	1623	QSGNLSR	1725	RSDALAR	1827	NRATLAR	1929	67.5
652	TAGGAGGCTG	1624	QSSDLQR	1726	RSDNLAR	1828	RSDNLTT	1930	1.5
653	TAGGAGGCTG	1625	TTSDLTR	1727	RSDNLAR	1829	RSDNLTT	1931	5.5
654	TAGGCATAAA	1626	QSGNLRT	1728	QSGSLTR	1830	RSDNLTT	1932	105
655	TAGGCATAAA	1627	QSGNLRT	1729	QSSTLRR	1831	RSDNLTT	1933	1000
656	TAGGCATAAA	1628	QSGNLRT	1730	QSGSLTR	1832	RSDNLTS	1934	540
657	TAGGCATAAA	1629	QSGNLRT	1731	QSSTLRR	1833	RSDNLTS	1935	300
660	GAGGGAGTTC	1630	NRATLAR	1732	QSGHLTR	1834	RSDNLAR	1936	8.25

T002T" 466666

661	GAGGGAGTTC	1631	TTSALTR	1733	QSGHLTR	1835	RSDNLAR	1937	1.73
665	GCGGAGGCGC	1632	RSDDVTR	1734	RSDNLTR	1836	RSDDLRR	1938	12.5
689	AAGGCGGAGA	1633	RSDNLTR	1735	RSDELQR	1837	RLDNRTA	1939	82.5
692	AAGGCGGAGA	1634	RSDNLTR	1736	RSDELQR	1838	RSDNLQ	1940	51
693	AAGGCGGAGA	1635	RSDNLTR	1737	RADTLRR	1839	RLDNRTA	1941	95
694	AAGGCGGAGA	1636	RSDNLTR	1738	RADTLRR	1840	RSDNLQ	1942	28.5
695	GGGGGCGAGC	1637	RSSNLTR	1739	DRSHLAR	1841	RSDHLTR	1943	850
697	TGAGCGGCGG	1638	RSDELTR	1740	RSDELSR	1842	QSGHLTK	1944	200
698	TGAGCGGCGG	1639	RSDELTR	1741	RSDELSR	1843	QSHGLTS	1945	300
699	GCGGCGGCAG	1640	QSGSLTR	1742	RSDDLQR	1844	RSDERKR	1946	21.5
700	GCGGCGGCAG	1641	QSGDLTR	1743	RSDDLQR	1845	RSDERKR	1947	45
701	GCAGCGGAGC	1642	RSDNLAR	1744	RSDELQR	1846	QSGSLTR	1948	50.5
702	GCAGCGGAGC	1643	RSDNLAR	1745	RSDELQR	1847	QSGDLTR	1949	73.5
704	AAGGTGGCAG	1644	QSGDLTR	1746	RSDSLAR	1848	RSDNLQ	1950	5
705	GGGGTGGGGC	1645	RSDHLAR	1747	RSDSLAR	1849	RSDHLSR	1951	0.01
706	GGGGTGGGGC	1646	RSDHLAR	1748	RSDSLLR	1850	RSDHLSR	1952	0.05
708	GAGTCGGAA	1647	QSANLAR	1749	RQDTLVG	1851	RSDNLAR	1953	300
709	GAGTCGGAA	1648	QSANLAR	1750	RKDVLVS	1852	RSDNLAR	1954	400
710	GAGTCGGAA	1649	QSGNLAR	1751	RLDGLRT	1853	RSDNLAR	1955	400
711	GAGTCGGAA	1650	QSGNLAR	1752	RQDTLVG	1854	RSDNLAR	1956	400
712	GGTGAGGAGT	1651	RSDNLAR	1753	RSDNLAR	1855	MSDHLSR	1957	9.5
713	GGTGAGGAGT	1652	RSDNLAR	1754	RSDNLAR	1856	MSHHLSR	1958	0.15
714	TGGGTCGCGG	1653	RSDELRR	1755	DRSALAR	1857	RSDHLTT	1959	200
715	TGGGTCGCGG	1654	RADTLRR	1756	DRSALAR	1858	RSDHLTT	1960	0.46
716	TTGGGAGCAC	1655	QSGSLTR	1757	QSGHLQR	1859	RGDALTS	1961	200
717	TTGGGAGCAC	1656	QSGSLTR	1758	QSGHLQR	1860	RSDALTK	1962	150
718	TTGGGAGCAC	1657	QSGSLTR	1759	QSGHLQR	1861	RSDALTR	1963	107.5
719	GGCATGGTGG	1658	RSDALTR	1760	RSDALTS	1862	DRSHLAR	1964	20
720	GAAGAGGATG	1659	TTSNLAR	1761	RSDNLAR	1863	QSGNLTR	1965	1.6
722	ATGGGGGTGG	1660	RSDALTR	1762	RSDHLTR	1864	RSDALRQ	1966	0.7
724	GGCATGGTGG	1661	RSDALTR	1763	RSDALRQ	1865	DRSHLAR	1967	2.5

099994 4666660

762	GCGGTGGCTC	1693	QSSDLQR	1795	RSDALSR	1897	RSDTLKK	1999	18
776	ATGGACGGGT	1694	RSDHLAR	1796	DRSNLER	1898	RSDSLNQ	2000	0.4
777	ATGGACGGGT	1695	RSDHLAR	1797	DRSNLTR	1899	RSDALSA	2001	3.4
779	CGGGGAGCAG	1696	QSGSLTR	1798	QSGHLTR	1900	RSDHLAE	2002	0.5
780	CGGGGAGCAG	1697	QSGSLTR	1799	QSGHLTR	1901	RSDHLRA	2003	0.5
781	GGGGAGCAGC	1698	RSSNLRE	1800	RSDNLAR	1902	RSDHLTR	2004	4.25
783	TTGGGAGCGG	1699	RSDELTR	1801	QSGHLQR	1903	RGDALTS	2005	2000
785	TTGGGAGCGG	1700	RSDTLKK	1802	QSGHLQR	1904	RSDALTS	2006	50
786	TTGGGAGCGG	1701	RSDTLKK	1803	QSGHLQR	1905	RGDALRS	2007	2000
787	AGGGAGGATG	1702	QSDNLAR	1804	RSDNLAR	1906	RSDHLTQ	2008	4
826	GAGGGAGCGA	1703	RSDELTR	1805	QSGHLAR	1907	RSDNLAR	2009	2.75
827	GAGGGAGCGA	1704	RADTLRR	1806	QSGHLAR	1908	RSDNLAR	2010	1.2
882	GCGTGGGCGT	1705	RSDELTR	1807	RSDHLTT	1909	RSDEKRR	2011	0.01
883	GCGTGGGCGT	1706	RSDELTR	1808	RSDHLTT	1910	RSDEKRR	2012	1

098994.1201

965	GCGGCGGCC	2053	ERGDLTR	2553	RSDELQR	3053	RSDEKR	3553	3.1
966	GAGGAGGCC	2054	ERGTLLAR	2554	RSDNLSR	3054	RSDNLAR	3554	0.028
967	GAGGAGGCC	2055	DRSSLTR	2555	RSDNLSR	3055	RSDNLAR	3555	0.055
968	GAGGCCGCA	2056	QSGSLTR	2556	DRSSLTR	3056	RSDNLAR	3556	1.4
969	GAGGCCGCA	2057	QSGSLTR	2557	DRSDLTR	3057	RSDNLAR	3557	0.275
970	GTGGGCGCC	2058	ERGTLLAR	2558	DRSHLAR	3058	RSDALAR	3558	1.859
971	GTGGGCGCC	2059	DRSSLTR	2559	DRSHLAR	3059	RSDALAR	3559	0.144
972	GTGGGCGCC	2060	ERGDLTR	2560	DRSHLAR	3060	RSDALAR	3560	1.748
973	GCCGCGGTC	2061	DRSALTR	2561	RSDELQR	3061	ERGTLLAR	3561	0.6
974	GCCGCGGTC	2062	DRSALTR	2562	RSDELQR	3062	DRSDLTR	3562	0.038
975	CAGGCCGCT	2063	QSSDLTR	2563	DRSSLTR	3063	RSDNLRE	3563	1.1
976	CAGGCCGCT	2064	QSSDLTR	2564	DRSDLTR	3064	RSDNLRE	3564	4.12
977	CTGGCAGTG	2065	RSDSLTR	2565	QSGSLTR	3065	RSDALRE	3565	0.017
978	CTGGCAGTG	2066	RSDSLTR	2566	QSGDLTR	3066	RSDALRE	3566	1.576
979	CTGGCGGCG	2067	RSSDLTR	2567	RSDELQR	3067	RSDALRE	3567	1.59
980	CTGGCGGCG	2068	RSDDLTR	2568	RSDELQR	3068	RSDALRE	3568	2.2
981	CAGGCGGCG	2069	RSDDLTR	2569	RSDELQR	3069	RSDNLRE	3569	0.375
982	CCGGGCTGG	2070	RSDHLTT	2570	DRSHLAR	3070	RSDELRE	3570	0.03
983	CCGGGCTGG	2071	RSDHLTK	2571	DRSHLAR	3071	RSDELRE	3571	1.385
984	GACGGCGAG	2072	RSDNLAR	2572	DRSHLAR	3072	DRSNLTR	3572	1.6
985	GACGGCGAG	2073	RSDNLAR	2573	DRSHLAR	3073	EKANLTR	3573	0.965
986	GGTGCTGAT	2074	QSSNLQR	2574	QSSDLQR	3074	MSHHLSR	3574	1.6
987	GGTGCTGAT	2075	QSSNLQR	2575	QSSDLQR	3075	TSGHLVR	3575	33.55
988	GGTGCTGAT	2076	TSGNLVR	2576	QSSDLQR	3076	MSHHLSR	3576	0.15
989	GGTGAGGGG	2077	RSDHLAR	2577	RSDNLAR	3077	MSHHLSR	3577	1.9
990	AAGGTGGGC	2078	DRSHLTR	2578	RSDSLAR	3078	RSDNLQ	3578	5.35
991	AAGGTGGGC	2079	DRSHLTR	2579	SSGSLVR	3079	RSDNLQ	3579	0.06
993	GGGGCTGGG	2080	RSDHLAR	2580	TSGELVR	3080	RSDHLAR	3580	3.1
994	GGGGCTGGG	2081	RSDHLTK	2581	DRSHLAR	3081	RSDHLAR	3581	0.03
995	GGGGAGGAA	2082	QSANLAR	2582	RSDNLAR	3082	RSDHLAR	3582	0.08
996	CAGTTGGTC	2083	DRSALAR	2583	RSDALTS	3083	RSDNLRE	3583	9.6
997	AGAGAGGCT	2084	QSSDLTR	2584	RSDNLAR	3084	QSGHLNQ	3584	1.65
998	ACGTAGTAG	2085	RSANLRT	2585	RSDNLTK	3085	RSDTLKQ	3585	0.23
999	AGAGAGGCT	2086	QSSDLTR	2586	RSDNLAR	3086	QSGKLTQ	3586	0.6
1000	CAGTTGGTC	2087	DRSALAR	2587	RSDALTR	3087	RSDNLRE	3587	11.15
1001	GGAGCTGAC	2088	EKANLTR	2588	QSSDLAR	3088	QRAHLAR	3588	1.8
1002	GCGGAGGAG	2089	RSDNLVR	2589	RSDNLAR	3089	RSDEKR	3589	0.028
1003	ACGTAGTAG	2090	RSANLRT	2590	RSDNLTK	3090	RSDTLRS	3590	0.118
1004	ACGTAGTAG	2091	RSDNLTT	2591	RSDNLTK	3091	RSDTLRS	3591	1.4
1006	GAGGGGCGG	2092	RSDDLTR	2592	RSDHLTR	3092	QRAHLAR	3592	0.898
1007	GAGAGAGAT	2093	QSSNLQR	2593	QSGHLTR	3093	RLHNLAR	3593	167
1008	GAGATGGAG	2094	RSDNLSR	2594	RSDSLTQ	3094	RLHNLAR	3594	0.4
1009	GAGATGGAG	2095	RSDNLSR	2595	RSDSLTQ	3095	RSDNLSR	3595	1.9
1010	GAGAGAGAT	2096	QSSNLQR	2596	QSGHLTR	3096	RSDNLAR	3596	8.2
1011	TTGGTGGCG	2097	RSADLTR	2597	RSDSLAR	3097	RSDSLTK	3597	0.03
1012	GACGTAGGG	2098	RSDHLTR	2598	QSSSLVR	3098	DRSNLTR	3598	0.032
1013	GAGAGAGAT	2099	QSSNLQR	2599	QSGHLNQ	3099	RSDNLAR	3599	0.15

[illegible]

[illegible]

1173	GCTGAAGGG	2241	RSDHLSR	2741	QSGNLAR	3241	QSSDLRR	3741	0.008
1174	GCTGAAGGG	2242	RSDHLSR	2742	QSSNLVR	3242	QSSDLRR	3742	0.018
1175	AAGGTCGCC	2243	DRSDLTR	2743	DPGALVR	3243	RSDNLTQ	3743	8.9
1176	GTGGGAGCC	2244	DRSDLTR	2744	QRAHLER	3244	RSDALTR	3744	4.1
1177	CCGGGCGCA	2245	QSGSLTR	2745	DRSHLAR	3245	RSDTLRE	3745	4.1
1178	GAGGATGGC	2246	DRSHLAR	2746	TSGNLVR	3246	RSDNLAR	3746	0.085
1179	GCAGCGCAG	2247	RSSNLRE	2747	RSSDLTR	3247	QSGSLTR	3747	2.735
1180	AAGGAAAGA	2248	QSGHLNQ	2748	QSGNLAR	3248	RSDNLTQ	3748	4.825
1181	TTGGCTATG	2249	RSDALRQ	2749	TSGELVR	3249	RGDALTS	3749	8.2
1182	CAGGAAGGC	2250	DRSHLAR	2750	QSGNLAR	3250	RSDNLRE	3750	1.48
1183	CAGGAAGGC	2251	DRSHLAR	2751	QSSNLVR	3251	RSDNLRE	3751	1.935
1184	AAGGAAAGA	2252	KNWKLQA	2752	QSGNLAR	3252	RSDNLTQ	3752	2.785
1185	AAGGAAAGA	2253	KNWKLQA	2753	QSHNLAR	3253	RSDNLTQ	3753	5.25
1186	GCCGAGGTG	2254	RSDSLLR	2754	RSKNLQR	3254	ERGTLAR	3754	27.5
1187	CTGGTGGGC	2255	DRSHLAR	2755	RSDALTR	3255	RSDALRE	3755	0.006
1188	GTAGTATGG	2256	RSDHLTT	2756	QSSSLVR	3256	QRASLAR	3756	2.74
1189	ATGGTTGAA	2257	QSANLAR	2757	TSGALTR	3257	RSDALRQ	3757	1.51
1190	ATGGCAGTG	2258	RSDALTR	2758	QSGDLTR	3258	RSDSLNQ	3758	1.484
1191	ATGGCAGTG	2259	RSDALTR	2759	QSGSLTR	3259	RSDSLNQ	3759	5.325
1192	ATGGCAGTG	2260	RSDALTR	2760	QSGDLTR	3260	RSDALTQ	3760	2.364
1193	ATGGCAGTG	2261	RSDALTR	2761	QSGSLTR	3261	RSDALTQ	3761	3.125
1194	GAGAAGGTG	2262	RSDALTR	2762	RSDNRTA	3262	RSDNLTR	3762	2.19
1195	GAGAAGGTG	2263	RSDALTR	2763	RSDNRTA	3263	RSSNLTR	3763	2.8
1197	GAAGGTGCC	2264	ERGDLTR	2764	MSHHLR	3264	QSGNLTR	3764	14.8
1199	ATGGAGAAG	2265	RSDNRTA	2765	RSDNLTR	3265	RSDALTQ	3765	3.428
1200	ATGGAGAAG	2266	RSDNRTA	2766	RSSNLTR	3266	RSDALTQ	3766	16.87
1201	ATGGAGAAG	2267	RSDNRTA	2767	RSHNLTR	3267	RSDALTQ	3767	14.8
1202	CTGGAGTAC	2268	DRSNLRT	2768	RSDNLTR	3268	RSDALRE	3768	2.834
1203	GGAGTACTG	2269	RSDALRE	2769	QRSALAR	3269	QRAHLAR	3769	2.945
1204	GGAGTACTG	2270	RSDALRE	2770	QSSSLVR	3270	QRAHLAR	3770	4.38
1205	CGGGCAGCT	2271	QSSDLRR	2771	QSGDLTR	3271	RSDHLRE	3771	0.9
1206	GCGGGAGTT	2272	TTSALTR	2772	QRAHLER	3272	RSDERKR	3772	0.034
1207	CAGGCTGGA	2273	QRAHLER	2773	TSGELVR	3273	RSDNLRE	3773	0.45
1209	CCGGAAGCG	2274	RSDELTR	2774	QSSNLVR	3274	RSDTLRE	3774	19.28
1211	GCAGCGCAG	2275	RSDNLRE	2775	RSDELTR	3275	QSGSLTR	3775	6.5
1212	CAGGGGGTT	2276	TTSALTR	2776	RSDHLTR	3276	RSDNLRE	3776	0.05
1213	GAAGAAGAG	2277	RSDNLTR	2777	QSSNLVR	3277	QSGNLTR	3777	12.3
1214	ATGGGAGTT	2278	TTSALTR	2778	QRAHLER	3278	RSDALTQ	3778	0.46
1215	GTGGGGGCT	2279	QSSDLRR	2779	RSDHLTR	3279	RSDALTR	3779	0.003
1217	GAAGAGGCA	2280	QSGSLTR	2780	RSDNLTR	3280	QSANLTR	3780	0.004
1218	GCGGTGAGG	2281	RSDHLTQ	2781	RSQALTR	3281	RSDERKR	3781	0.46
1219	AAGGAAAGG	2282	RSDHLTQ	2782	QSHNLAR	3282	RSDNLTQ	3782	0.68
1220	AAGGAAAGG	2283	RSDHLTQ	2783	QSGNLAR	3283	RSDNLTQ	3783	0.175
1221	AAGGAAAGG	2284	RSDHLTQ	2784	QSSNLVR	3284	RSDNLTQ	3784	1.4
1222	CAGGAGGGC	2285	DRSHLAR	2785	RSDNLAR	3285	RSDNLRE	3785	0.155
1223	ATGGACTTG	2286	RSDALTK	2786	DRSNLTR	3286	RSDALTQ	3786	7
1224	ATGGACTTG	2287	RADALMV	2787	DRSNLTR	3287	RSDALTQ	3787	12

F002T"4666666

1227	GAATAGGGG	2288	RSDHLSR 2788	RSDHLTK 3288	QSGNLAR 3788	25
1228	ACGGCCGAG	2289	RSDNLAR 2789	DRSDLTR 3289	RSDDLQ 3789	12
1229	AAGGGTGCG	2290	RSDELTR 2790	MSHLSR 3290	RSDNLQ 3790	8.2
1230	AAGGGAGAC	2291	DRSNLTR 2791	QSGHLTR 3291	RSDNLQ 3791	0.383
1231	AAGGGAGAC	2292	DRSNLTR 2792	QRAHLER 3292	RSDNLQ 3792	0.213
1232	TGGGACCTG	2293	RSDALRE 2793	DRSNLTR 3293	RSDHLTT 3793	0.113
1233	TGGGACCTG	2294	RSDALRE 2794	DRSNLTR 3294	RSDHLTT 3794	0.635
1234	GAGTAGGCA	2295	QSGSLTR 2795	RSDNLTK 3295	RSDNLAR 3795	0.101
1236	GAGTAGGCA	2296	QSGSLTR 2796	RSDHLTT 3296	RSDNLAR 3796	0.065
1237	GAAGGAGAG	2297	RSDNLAR 2797	QRAHLER 3297	QSGNLAR 3797	0.065
1238	CTGGATGTT	2298	QSSALAR 2798	TSGNLVR 3298	RSDALRE 3798	0.313
1239	CAGGACGTG	2299	RSDALTR 2799	DPGNLVR 3299	RSDNLKD 3799	0.144
1240	GGGGAGGCA	2300	QSGSLTR 2800	RSDNLTR 3300	RSDHLSR 3800	0.056
1241	GAGGTGTCA	2301	QSHDLTK 2801	RSDALAR 3301	RSDNLAR 3801	0.027
1242	GGGGTTGAA	2302	QSANLAR 2802	TSGSLTR 3302	RSDHLSR 3802	0.02
1243	GGGGTTGAA	2303	QSANLAR 2803	QSSALTR 3303	RSDHLSR 3803	0.101
1244	GTCGCGGTG	2304	RSDALTR 2804	RSDELQR 3304	DRSALAR 3804	0.044
1245	GTCGCGGTG	2305	RSDALTR 2805	RSDELQR 3305	DSGSLTR 3805	0.102
1246	GTGGTTGCG	2306	RSDELTR 2806	TSGSLTR 3306	RSDALTR 3806	0.051
1247	GTGGTTGCG	2307	RSDELTR 2807	TSGALTR 3307	RSDALTR 3807	0.117
1248	GTCTAGGTA	2308	QSGALTR 2808	RSDNLTT 3308	DRSALAR 3808	5.14
1249	CCGGGAGCG	2309	RSDELTR 2809	QSGHLTR 3309	RSDTLRE 3809	0.26
1250	GAAGGAGAG	2310	RSDNLAR 2810	QSGHLTR 3310	QSGNLAR 3810	0.31
1252	CCGGCTGGA	2311	QRAHLER 2811	QSSDLTR 3311	RSDTLRE 3811	0.153
1253	CCGGGAGCG	2312	RSDELTR 2812	QRAHLER 3312	RSDTLRE 3812	0.228
1255	ACGTAGTAG	2313	RSDNLTT 2813	RSDNLTK 3313	RSDTLKQ 3813	0.69
1256	GGGGAGGAT	2314	QSSNLAR 2814	RSDNLQR 3314	RSDHLSR 3814	2
1257	GGGGAGGAT	2315	TTSNLAR 2815	RSDNLQR 3315	RSDHLSR 3815	1
1258	GGGGAGGAT	2316	QSSNLRR 2816	RSDNLQR 3316	RSDHLSR 3816	2
1259	GAGTGTGTG	2317	RSDSLLR 2817	DRDHLTR 3317	RSDNLAR 3817	1.5
1260	GAGTGTGTG	2318	RLDSL LR 2818	DRDHLTR 3318	RSDNLAR 3818	1.8
1261	TGCGGGGCA	2319	QSGDLTR 2819	RSDHLTR 3319	RRDTLHR 3819	0.2
1262	TGCGGGGCA	2320	QSGDLTR 2820	RSDHLTR 3320	RLDTLGR 3820	3
1263	TGCGGGGCA	2321	QSGDLTR 2821	RSDHLTR 3321	DSGHLAS 3821	21
1264	AAGTTGGTT	2322	TTSALTR 2822	RADALMV 3322	RSDNLQ 3822	0.21
1265	AAGTTGGTT	2323	TTSALTR 2823	RSDALTT 3323	RSDNLQ 3823	0.077
1266	CAGGGTGGC	2324	DRSHLTR 2824	QSSHLAR 3324	RSDNLRE 3824	6.1
1267	TAGGCAGTC	2325	DRSALTR 2825	QSGSLTR 3325	RSDNLTT 3825	6
1268	CTGTTGGCT	2326	QSSDLTR 2826	RADALMV 3326	RSDALRE 3826	1.52
1269	CTGTTGGCT	2327	QSSDLTR 2827	RSDALTT 3327	RSDALRE 3827	12.3
1270	TTGGATGGA	2328	QSGHLAR 2828	TSGNLVR 3328	RSDALTK 3828	0.4
1271	GTGGCACTG	2329	RSDALRE 2829	QSGSLTR 3329	RSDALTR 3829	0.915
1272	CAGGAGTCC	2330	DRSSLTT 2830	RSDNLAR 3330	RSDNLRE 3830	0.04
1273	CAGGAGTCC	2331	ERGD LTT 2831	RSDNLAR 3331	RSDNLRE 3831	0.1
1274	GCATGGGAA	2332	QSANLSR 2832	RSDHLTT 3332	QSGSLTR 3832	0.306
1275	GCATGGGAA	2333	QRSNLVR 2833	RSDHLTT 3333	QSGSLTR 3833	0.326
1276	TAGGAAGAG	2334	RSDNLAR 2834	QRSNLVR 3334	RSDNLTT 3834	0.685

TOTAL " 4668650

1277	GAAGAGGGG	2335	RSDHLAR	2835	RSDNLAR	3335	QSGNLTR	3835	0.421
1278	GAGTAGGCA	2336	QSGSLTR	2836	RSDNLRT	3336	RSDNLAR	3836	0.019
1279	GAGGTGTCA	2337	QSGDLRT	2837	RSDALAR	3337	RSDNLAR	3837	0.025
1282	TCGGTCGCC	2338	ERGDLTR	2838	DPGALVR	3338	RSDELRT	3838	74.1
1287	GTGGTAGGA	2339	QSGHLAR	2839	QSGALAR	3339	RSDALTR	3839	0.152
1288	CAGGGTGGC	2340	DRSHLTR	2840	QSSHLAR	3340	RSDNLTE	3840	4.1
1289	TAGGCAGTC	2341	DRSALTR	2841	QSGSLTR	3341	RSDNLTK	3841	1.37
1290	GTGGTGATA	2342	QSGALTQ	2842	RSHALTR	3342	RSDALTR	3842	24.05
1291	GTGGTGATA	2343	QQASLNA	2843	RSHALTR	3343	RSDALTR	3843	20.55
1292	TTGGATGGA	2344	QSGHLAR	2844	TSGNLVR	3344	RSDALTT	3844	4.12
1293	AAGGTAGGT	2345	TSGHLVR	2845	QSGALAR	3345	RSDNLTK	3845	0.457
1294	AAGGTAGGT	2346	MSHHLSR	2846	QSGALAR	3346	RSDNLTK	3846	2.75
1295	CAGGAGTCC	2347	DRSSLTT	2847	RSDNLAR	3347	RSDNLTE	3847	0.116
1296	CAGGAGTCC	2348	ERGDLT	2848	RSDNLAR	3348	RSDNLTE	3848	37
1297	TAGGAAGAG	2349	RSDNLAR	2849	QRSNLVR	3349	RSDNLTK	3849	0.05
1298	CAGGACGTG	2350	RSDLATR	2850	DPGNLVR	3350	RSDNLTE	3850	0.05
1300	GTCTAGGTA	2351	QSGALTR	2851	RSDNLTK	3351	DRSALAR	3851	0.46
1302	CCGGCTGGA	2352	QSGHLTR	2852	QSSDLTR	3352	RSDTLRE	3852	0.05
1303	TAGGAGTTT	2353	QRSALAS	2853	RSDNLAR	3353	RSDNLTK	3853	0.088
1306	CTGGCCTTG	2354	RSDALTT	2854	DCRDLAR	3354	RSDALRE	3854	2.285
1308	TGGGCAGCC	2355	ERGTLAR	2855	QSGSLTR	3355	RSDHLTT	3855	0.305
1309	TAGGAGTTT	2356	QSSALAS	2856	RSDNLAR	3356	RSDNLTK	3856	0.184
1310	TAGGAGTTT	2357	TTSALAS	2857	RSDNLAR	3357	RSDNLTK	3857	0.075
1311	TGGGCAGCC	2358	ERGDLAR	2858	QSGSLTR	3358	RSDHLTT	3858	0.91
1312	GGGGCGTGA	2359	QSGHLTK	2859	RSDELQR	3359	RSDHLSR	3859	0.23
1313	GGGGCGTGA	2360	QSGHLTT	2860	RSDELQR	3360	RSDHLSR	3860	0.09
1314	GTACAGTAG	2361	RSDNLTK	2861	RSDNLRE	3361	QSSSLVR	3861	3.09
1315	GTACAGTAG	2362	RSDNLTK	2862	RSDNLTE	3362	QSSSLVR	3862	9.27
1318	ATGGTGTGT	2363	TSSHLLS	2863	RSDALAR	3363	RSDALAQ	3863	0.048
1319	ATGGTGTGT	2364	MSHHLTT	2864	RSDALAR	3364	RSDALAQ	3864	0.228
1320	TTGGGAGAG	2365	RSDNLAR	2865	QRAHLER	3365	RSDALTT	3865	0.044
1321	TTGGGAGAG	2366	RSDNLAR	2866	QRAHLER	3366	RADALMV	3866	0.127
1322	GTGGGAATA	2367	QSGALTQ	2867	QSGHLTR	3367	RSDALTR	3867	0.799
1323	GTGGGAATA	2368	QLTGLNQ	2868	QSGHLTR	3368	RSDALTR	3868	0.744
1324	GTGGGAATA	2369	QQASLNA	2869	QSHHLTR	3369	RSDALTR	3869	18.52
1325	TTGGTTGGT	2370	TSGHLVR	2870	TSGSLTR	3370	RSDALTK	3870	0.306
1326	TTGGTTGGT	2371	TSGHLVR	2871	QSSALTR	3371	RSDALTK	3871	4.385
1327	TTGGTTGGT	2372	TSGHLVR	2872	TSGSLTR	3372	RSDALTT	3872	0.566
1328	TTGGTTGGT	2373	TSGHLVR	2873	QSSALTR	3373	RSDALTT	3873	7.95
1329	CTGGCCTGG	2374	RSDHLTT	2874	DRSDLTR	3374	RSDALRE	3874	0.68
1330	GAGGTGTGA	2375	QSGHLTT	2875	RSDALTR	3375	RSDNLAR	3875	0.175
1331	CTGGCCTGG	2376	RSDHLTT	2876	DCRDLAR	3376	RSDALRE	3876	0.388
1334	CCGGCGCTG	2377	RSDALRE	2877	RSSDLTR	3377	RSDDLRE	3877	0.31
1335	GACGCTGGC	2378	DRSHLTR	2878	QSSDLTR	3378	DSSNLTR	3878	1.4
1336	CGGGCTGGA	2379	QSGHLAR	2879	QSSDLTR	3379	RSDHLAE	3879	1.4
1337	CGGGCTGGA	2380	QSSHLAR	2880	QSSDLTR	3380	RSDHLAE	3880	0.235
1338	GGGATGGCG	2381	RSDELTR	2881	RSDALTQ	3381	RSDHLSR	3881	1.04

100211.20

[illegible]

FOOT " 4666660

1461	GACGAGGAG	2476	RSANLAR	2976	RSDNLTR	3476	DRSNLTR	3976	0.014
1462	CGGGATGAA	2477	QSGNLAR	2977	TSGNLVR	3477	RSDHLRE	3977	0.05
1463	GAGGCTGTT	2478	TTSALTR	2978	QSSDLTR	3478	RSDNLAR	3978	0.003
1464	GACGAGGAG	2479	RSDNLAR	2979	RSDNLTR	3479	DRSNLTR	3979	0.002
1465	CTGGGAGTT	2480	TTSALTR	2980	QSGHLQR	3480	RSDALRE	3980	0.018
1466	CTGGGAGTT	2481	NRATLAR	2981	QSGHLQR	3481	RSDALRE	3981	0.017
1468	GGTGATGTC	2482	DRSALTR	2982	TSGNLVR	3482	MSHHLR	3982	0.08
1469	GGTGATGTC	2483	DRSALTR	2983	TSGNLVR	3483	TSGHLVR	3983	0.28
1470	GGTGATGTC	2484	DRSALTR	2984	TSGNLVR	3484	QRAHLR	3984	0.156
1471	CTGGTTGGG	2485	RSDHLR	2985	QSSALTR	3485	RSDALRE	3985	0.09
1472	TTGAAGGTT	2486	TTSALTR	2986	RSDNLQ	3486	RADALMV	3986	3.22
1473	TTGAAGGTT	2487	TTSALTR	2987	RSDNLQ	3487	RSDSLTT	3987	0.47
1474	TTGAAGGTT	2488	QSSALAR	2988	RSDNLQ	3488	RADALMV	3988	1.39
1475	TTGAAGGTT	2489	QSSALAR	2989	RSDNLQ	3489	RLHSLTT	3989	0.39
1476	TTGAAGGTT	2490	QSSALAR	2990	RSDNLQ	3490	RSDSLTT	3990	0.305
1477	GCAGCCCGG	2491	RSDHLRE	2991	DRSDLTR	3491	QSGSLTR	3991	2.31
1479	GAAAGTTCA	2492	QSHDLTK	2992	MSHHLQ	3492	QSGNLAR	3992	37.04
1480	GAAAGTTCA	2493	NKTDLGK	2993	TSGHLVQ	3493	QSGNLAR	3993	62.5
1481	GAAAGTTCA	2494	NKTDLGK	2994	TSDHLAS	3494	RSDELRE	3994	37.04
1482	CCGTGTGAC	2495	DRSNLTR	2995	TSDHLAS	3495	RSDELRE	3995	111.1
1483	CCGTGTGAC	2496	DRSNLTR	2996	MSHHLTT	3496	RSDELRE	3996	20.8
1484	GAAGTGGTA	2497	QSSSLVR	2997	RSDALSR	3497	QSGNLAR	3997	0.01
1485	AAGTGAGCT	2498	QSSDLRR	2998	QSGHLTT	3498	RSDNLQ	3998	1.537
1486	GGGTTTGAC	2499	DRSNLTR	2999	TTSALAS	3499	RSDHLR	3999	0.085
1487	TTGAAGGTT	2500	TTSALTR	3000	RSDNLQ	3500	RLHSLTT	4000	0.188
1488	AAGTGGTAG	2501	QSSDLRR	3001	QSGHLTT	3501	RLDNRTQ	4001	5.64
1490	CTGGTTGGG	2502	RSDHLR	3002	TSGSLTR	3502	RSDALRE	4002	0.04
1491	AAGGGTTCA	2503	NKTDLGK	3003	DSSKLSR	3503	RLDNRTA	4003	4.12
1492	AAGTGGTAG	2504	RSDNLTT	3004	RSDHLTT	3504	RSDNLQ	4004	1.37
1493	AAGTGGTAG	2505	RSDNLTT	3005	RSDHLTT	3505	RLDNRTQ	4005	15.09
1494	GGGTTTGAC	2506	DRSNLTR	3006	QRSALAS	3506	RSDHLR	4006	0.255
1496	TTGGGGGAG	2507	RSDNLAR	3007	RSDHLTR	3507	RSDALTT	4007	0.065
1497	GAGGCTCTT	2508	QSSALAR	3008	QSSDLTR	3508	RSDNLAR	4008	0.007
1498	GAGGTTGAT	2509	QSSNLAR	3009	QSSALTR	3509	RSDNLAR	4009	0.101
1499	GAGGTTGAT	2510	QSSNLAR	3010	TSGALTR	3510	RSDNLAR	4010	0.02
1500	GCAGAGGAA	2511	QSGNLAR	3011	RSDNLAR	3511	QSGSLTR	4011	0.003
1522	GCAATGGGT	2512	TSGHLVR	3012	RSDALTQ	3512	QSGDLTR	4012	0.08

TABLE 6

TRIPLET (5'→3')	FINGER (N → C)		
	F1	F2	F3
AGG			RXDHXXQ
ATG			RXDAXXQ
CGG			RXDHXXE
GAA		QXGNXXR	
GAC	DXSNXXR		DXSNXXR
GAG	RXDNXXR	RXSNXXR RXDNXXR	RXDNXXR
GAT	QXSNXXR TXSNXXR TXGNXXR	TXGNXXR	
GCA	QXGSXXR	QXGDXXR	
GCC	EXGTXXR		
GCG	RXDEXXR	RXDEXXR	RXDEXXR RXDTXXK
GCT	QXSDXXR	TXGEXXR QXSDXXR	
GGA		QXGHXXR	QXAHXXR
GGC	DXSHXXR	DXSHXXR	
GGG	RXDHXXR	RXDHXXR	RXDHXXR RXDHXXK
GGT			TXGHXXR
GTA		QXGSXXR QXATXXR	
GTG	RXDAXXR RXDSXXR	RXDAXXR	RXDAXXR
TAG		RXDNXXT	
TCG	RXDDXXK		
TGT		TXDHXXS	

T002T-466860